Revolutionizing Science and Engineering Through Cyberinfrastructure:

Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure

January 2003

Daniel E. Atkins, Chair
University of Michigan

Kelvin K. Droegemeier
University of Oklahoma

Stuart I. Feldman
IBM

Hector Garcia-Molina
Stanford University

Michael L. Klein
University of Pennsylvania

David G. Messerschmitt
University of California at Berkeley

Paul Messina
California Institute of Technology

Jeremiah P. Ostriker
Princeton University

Margaret H. Wright
New York University

Disclaimer

This report was prepared by an officially appointed advisory panel to the National Science Foundation, however, any opinions, findings, and conclusions or recommendations expressed in this material are those of the panel and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

1.0 Introduction

Scientific and engineering research has been crucial in both the *creation* and the advanced *application* of the amazing products of the digital revolution begun some sixty years ago – a revolution that increasingly undergirds our modern world. Advances in computational technology continue to transform scientific and engineering research, practice, and allied education. Recently, multiple accelerating trends are converging and crossing thresholds in ways that show extraordinary promise for an even more profound and rapid transformation – indeed a further revolution – in how we create, disseminate, and preserve scientific and engineering knowledge. We now have the opportunity and responsibility to integrate and extend the products of the digital revolution to serve the next generation of science and engineering research and education.

Digital computation, data, information, and networks are now being used to replace and extend traditional efforts in science and engineering research, indeed to create new disciplines. The classic two approaches to scientific research, theoretical/analytical and experimental/observational, have been extended to *in silico* simulation to explore a larger number of possibilities at new levels of temporal and spatial fidelity. Advanced networking enables people, tools, and information to be linked in ways that reduce barriers of location, time, institution, and discipline. In numerous fields new distributed-knowledge environments are becoming essential, not optional, for moving to the next frontier of research. Science and engineering researchers are again at the forefront in both creating and exploiting what many are now seeing as a nascent revolution and a forerunner of new capabilities for broad adoption in our knowledge-driven society.

A vast opportunity exists for creating new research environments based upon cyberinfrastructure, but there are also real dangers of disappointing results and wasted investment for a variety of reasons including underfunding in amount and duration, lack of understanding of technological futures, excessively redundant activities between science fields or between science fields and industry, lack of appreciation of social/cultural barriers, lack of appropriate organizational structures, inadequate related educational activities, and increased technological ("not invented here") balkanizations rather than interoperability among multiple disciplines. The opportunity is enormous, but also enormously complex, and must be approached in a long-term, comprehensive way. It is imperative to begin a well-conceived and funded program to seize these opportunities and to avoid potentially increasing opportunity costs.

This report is from a Blue Ribbon Panel convened by the Assistant Director for Computer and Information Science and Engineering (CISE)[1]* of the National Science Foundation (NSF) to inventory and explore these trends and to make strategic recommendations on the nature and form of programs that NSF should take in response to them. The charge to the Panel is premised on the concept of an advanced infrastructure layer on which innovative science and engineering research and education environments can be built. The term *infrastructure* has been used since the 1920s to refer collectively to the roads, power grids, telephone systems, bridges, rail lines, and similar public works that are required for an industrial economy to function. Although good infrastructure is often taken for granted and noticed only when it stops functioning, it is among the most complex and expensive thing that society creates. The newer term *cyberinfrastructure* refers to infrastructure based upon distributed computer, information and communication technology. If *infrastructure* is required for an *industrial* economy, then we could say that *cyberinfrastructure* is required for a *knowledge* economy.

The charge to the Panel is to 1) evaluate current major investments in cyberinfrastructure, most especially the Partnerships for Advanced Computational Infrastructure (PACI)[2]; 2) recommend new areas of emphasis relevant to cyberinfrastructure; and 3) propose an implementation plan for pursuing these new areas of emphasis. The full text of the charge is included as Appendix E.

The base technologies underlying cyberinfrastructure are the integrated electro-optical components of computation, storage, and communication that continue to advance in raw capacity at exponential rates. Above the cyberinfrastructure layer are software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice. Between these two layers is the *cyberinfrastructure* layer of enabling hardware, algorithms, software, communications, institutions, and personnel. This layer should provide an effective and efficient platform for the empowerment of specific communities of researchers to innovate and eventually revolutionize what they do, how they do it, and who participates.

Although the term cyberinfrastructure is new, NSF investment in envisioning, creating, deploying, and using computational-based infrastructure is not. Previous NSF programs have created key capabilities and experience that have already done much to enable a next big step up in the power, ubiquity, and application of advanced cyberinfrastructure. They have been instrumental in creating the vision and demand for more. By *advanced* we mean both the highest-performing technology and its use in the most leading-edge research.

**Pointers to references are noted with superscripts and the citations are listed in Section 7.*

In the 1960s NSF funded some of the very first academic computing centers and in the 1970s funded early activities in computational science. Beginning in the mid 1980s the Advanced Scientific Computing (ASC) initiatives together with NSFNET provided the research community access to machines at the top of the computation pyramid. The NSFNET transitioned into the commercial Internet, and a decade later the ASC program evolved into a more comprehensive source of high-end computing and related services. Two Partnerships for Advanced Computing Infrastructure (PACI)[2] were formed: one centered at the National Center for Supercomputing Applications (NCSA)[3] at the University of Illinois, Urbana-Champaign, and the other at the San Diego Supercomputer Center (SDSC)[4] at the University of California, San Diego. Recently the NSF made awards for terascale capability facilities to the Pittsburgh Supercomputing Center (PSC)[5], and then awards for a Distributed Terascale Facility (tera*grid* capability)[6] to a project consortium including NCSA, SDSC, Argonne National Laboratory, the Center for Advanced Computing Research (CACR) at the California Institute of Technology, and the PSC.

The Terascale Initiative is providing network access to high-end computing through physically proximate clusters of commodity computation servers. The more recent Distributed Terascale Facility is continuing the exploration of new modes of computing by extending the concept of clusters to that of wide-area grids of supercomputers allocated dynamically to a common problem over both wide distance and multiple organizations.

Two other highly relevant initiatives are the NSF Middleware[7] and the Digital Library Initiatives[8]. The NSF Middleware Initiative and Integration Testbed is an ongoing effort to develop, disseminate, and evaluate software that allows scientists and educators easily to build and share new distributed applications, share instrumentation, and share access to common data repositories. The Digital Library Initiative has been a major catalyst in creating the vast information sources and new services of the Internet including Google. Likewise, basic research in computer and information science over many years has produced much of what we now know as the Internet and the Web.

The NSF CISE Directorate supported most of the initiatives cited above. But also emerging across all NSF directorates are a variety of multidisciplinary research communities, working in partnership with computer and information scientists and engineers, to explore how to revolutionize both what problems they explore, as well as how they go about exploring them. Generic names for such cyberinfrastructure-enabled environments include *collaboratory, co-laboratory, grid community/network, virtual science community,* and *e-science community.* Examples of specific science-driven pilot projects include the Network for Earthquake Engineering Simulation (NEES)[9], the National Virtual Observatory (NVO)[10], the National Ecological Observatory Network (NEON)[11], the National Science Digital Library

(NSDL)[12], the Grid Physics Network (GriPhyN)[13], and the Space Physics and Aeronomy Research Collaboratory (SPARC)[14]. Taken together with the CISE-based activities, these new projects are *building out* in terms of broader scientific application, and they are *building up* in terms of function and performance. They provide a glimpse into an exciting future.

Mission-oriented research agencies are also initiating similar projects, for example the NIH Biomedical Informatics Research Network (BIRN)[15], the Department of Energy (DOE) National Collaboratories Program[16], and the DoE project for Scientific Discovery Through Advanced Computing (SciDAC)[17]. Relevant international programs include the UK E-science program[18], parts of the EU 6th Framework Project[19], and the Japanese Earth Simulator Center[20].

As indicated by the title of this report, the scope of our exploration and recommendations goes well beyond the topic of cyberinfrastructure in isolation or as an end in itself. Building, operating, and using advanced cyberinfrastructure must be done in a systemic context that exploits mutual self-interest and synergy among computer and information, and social science research communities who see it as an *object of research*, and other ("domain science") research communities who see it as a platform in *service of research*. More specifically, we need highly coordinated, large, and long-term investment in

1. *fundamental research* to advance cyberinfrastructure;
2. *development activities* to create and evolve the building blocks of advanced operational cyberinfrastructure;
3. institutions with people and facilities to provide *operational support and services*; and
4. *high-impact applications* of advanced cyberinfrastructure in all areas of science and engineering research and allied education.

We envision the creation of thousands of overlapping field and project specific collaboratories or grid communities, customized at the application layer but extensively sharing common cyberinfrastructure. The cyberinfrastructure should include grids of computational centers, some with computing power second to none; comprehensive libraries of digital objects including programs and literature; multidisciplinary, well-curated federated collections of scientific data; thousands of online instruments and vast sensor arrays; convenient software toolkits for resource discovery, modeling, and interactive visualization; and the ability to collaborate with physically distributed teams of people using all of these capabilities. This vision requires enduring institutions with highly competent professionals to create and procure robust software, leading-edge hardware, specialized instruments, knowledge management facilities, and appropriate training.

Furthermore, cutting across all these coordinated endeavors we need specific activities to benefit education, general science awareness, and policymaking. We need coordinated participation by academia, private industry, non-NSF government agencies and laboratories, and state, regional, and national centers. A program in this area should be interagency and international. It must address very complex interaction between scientific, technological, and sociological challenges and opportunities.

The Panel's findings and recommendations have been informed by extensive interaction with broad areas of the scientific and engineering research communities through 62 presentations at invitational public testimony sessions (see Appendix D); 700 responses to a community-wide survey (see Appendix B); review of dozens of prior relevant reports; scores of unsolicited emails and phone calls; 250 pages of written critique from 60 reviewers of an early draft of this report; panel members attending conferences and workshops concerning visions and needs of specific research communities; and hundreds of hours of deliberation and discussion among Panel members. The members of the Panel have backgrounds in areas widely relevant to creating, managing, and using advanced cyberinfrastructure. They include high-performance computing, visualization, technology trends, digital libraries, databases, distributed systems, middleware, and collaboration technology. Members of the Panel also have considerable collective experience in industrial management and academic administration.

In the next section of this report we present our vision for an Advanced Cyberinfrastructure Program that we recommend be initiated immediately under the leadership of the NSF. We next summarize trends and issues that we believe are converging to motivate, justify, enable, and to some extent prescribe the Advanced Cyberinfrastructure Program we described in Section 2. In the remaining sections we discuss the principal requirements for achieving this program, primarily organizational and financial. We also discuss the role of the current major centers and projects now providing advanced cyberinfrastructure, particularly, as we were specifically asked, the PACI programs. Section 7. contains references. Supplementary material is included in five appendixes.

2.0 Vision for an Advanced Cyberinfrastructure Program

2.1 A Nascent Revolution

Scientists in many disciplines have begun revolutionizing their fields by using computers, digital data, and networks to replace and extend their traditional efforts. The calculations that can be performed and the information that can be archived and used are exploding. In the not-too-distant future, the contents of the historic scientific literature will fit on a rack of disks, and an office computer will provide more computing than all the supercomputing centers together today. The results of today's largest calculations and most sizable collections will take seconds to transmit using the fastest known network technologies. New technology-mediated, distributed work environments are emerging to relax constraints of distance and time. These new research environments are linking together research teams, digital data and information libraries, high-performance computational services, scientific instruments, and arrays of sensors. In many cases these emerging environments for knowledge work are essential, not optional, to the aspirations of research. We see glimpses of the future in some shifts in current research practice:

- The classic two approaches to scientific research, theoretical/analytical and experimental/observational, have been extended to *in silico* simulation and modeling to explore new possibilities and to achieve new precision.
- The enormous speedups of computers and networks have enabled simulations of far more complex systems and phenomena, as well as visualizing the results from many perspectives.
- Advanced computing is no longer restricted to a few research groups in a few fields such as weather prediction and high-energy physics, but pervades scientific and engineering research, including the biological, chemical, social, and environmental sciences, medicine, and nanotechnology.
- The primary access to the latest findings in a growing number of fields is through the Web, then through classic preprints and conferences, and lastly through refereed archival papers.
- Crucial data collections in the social, biological, and physical sciences are now online and remotely accessible – modern genome research would be impossible without such databases, and soon astronomical research will be similarly redefined through the National Virtual Observatory.
- Groups collaborate across institutions and time zones, sharing data, complementary expertise, ideas, and access to special facilities without travel.

The trends represented by these examples will only accelerate. In the future, we might expect researchers to

- Combine raw data and new models from many sources, and utilize the most up-to-date tools to analyze, visualize, and simulate complex interrelations.
- Collect and make widely available far more information (the outputs of all major observatories and astronomical satellites, satellite and land-based weather data, three-dimensional images of anthropologically important objects), leading to a qualitative change in the way research is done and the type of science that results.
- Work across traditional disciplinary boundaries: environmental scientists will take advantage of climate models, physicists will make direct use of astronomical observations, social scientists will analyze interactive behavior of scientists as well as others.
- Simulate more complex and exciting systems (cells and organisms rather than proteins and DNA; the entire earth system rather than air, water, land, and snow independently).
- Access the entire published record of science online.
- Make publications incorporating rich media (hypertext, video, photographic images).
- Visualize the results of complex data sets in new and exciting ways, and create techniques for understanding and acting on these observations.
- Work routinely with colleagues at distant institutions, even ones that are not traditionally considered research universities, and with junior scientists and students as genuine peers, despite differences in age, experience, race, or physical limitations.

2.2 Thresholds and Opportunities

Why act now? Currently observed activities and benefits represent just the beginnings of a revolution. Computers have been improving for decades, and some researchers have tried to do many of the activities listed above. We believe that several key thresholds have recently been reached in the use of IT, in part because NSF has made large and successful investments in a number of research areas, including networking, supercomputing, human interfaces, collaboration environments, and information management. There are many reasons:

- The Internet and the Web were invented to support the work of researchers, and their use permeates all of science and engineering. Broadband networks connect all research centers and enable the rapid communication of ideas, the sharing of resources, and remote access to data. The next generations of the net promise even greater benefits to the research community.
- Most modern researchers are fully conversant with and dependent on advanced computing for their daily activity, and have a thirst for more. Older scientists are learning to take advantage of the new technologies.

- Closed-form analytic solutions are available for a decreasing fraction of interesting research challenges; often only a numeric computation can produce useful results.
- Moore's law has led to simulations that begin to match the complexity of the real world, with fully three dimensional, time-dependent modeling with realistic physical models opening up a vast range of problems to qualitative attacks. They range from cosmology to protein folding – problems formerly considered far too complex to address directly.
- In an increasing fraction of cases, it is faster, cheaper, and more accurate to simulate a model than construct and observe a physical object.
- Increasingly ubiquitous networking and interoperability of information formats and access make high-quality remote collaboration feasible.
- Storing terabytes of information is common and inexpensive; archives containing hundreds or thousands of terabytes of data will be affordable and necessary for archiving scientific and engineering information.
- Computing power that was unavailable only a few years ago – trillions of operations a seconds – can now be found in a number of research organizations.
- Computational and visualization techniques have progressed enormously and provide as much scientific value as improved hardware.
- Most researchers would not be able to function without e-mail or access to the Web. They certainly would have fewer contacts with distant, especially international, scientists and be much less able to stay on the cutting edge of their field.

There are also significant risks and costs if we do not make a major move at this time:

- Absent coordination, researchers in different fields and at different sites will adopt different formats and representations of key information, which will make it forever difficult or impossible to combine or reconcile.
- Absent systematic archiving and curation of intermediate research results (as well as the polished and reduced publications), data gathered at great expense will be lost.
- Effective use of cyberinfrastructure can break down artificial disciplinary boundaries, while incompatible tools and structures can isolate scientific communities for years.
- Groups are building their own application and middleware software without awareness of comparable needs elsewhere, both within the NSF and across all of science. Much of this software will be of limited long-term value absent a consistent computer science perspective. Time and talent will be wasted that could have led to much better computing *and* much better science.

- Dramatic changes are coming in computing and application architectures; lack of consideration of work in other sciences and in the commercial world could render projects obsolete before they deliver.
- Much of the effort under way to use cyberinfrastructure for collaborative research is not giving adequate attention to sociological and culture barriers to technology adoption that may cause failure, even after large investments.

The time is ripe for NSF to accelerate the revolution for the benefit of society. A confluence of technology-push and science and engineering research-pull activities and possibilities makes this the right time. Researchers are ramping up their use of computing resources, starting to store enormous amounts of information, and sharing it. Distributed computing, large clusters, data farms, and broadband networks (typified by Internet2 [21], Grid[22], and Web Services[23] directions) have moved from research to practical use. We anticipate a phase change, where direct attention to this opportunity can have a highly desirable and nonlinear effect.

We envision an environment in which raw data and recent results are easily shared, not just within a research group or institution but also between scientific disciplines and locations. There is an exciting opportunity to share insights, software, and knowledge, to reduce wasteful re-creation and repetition. Key applications and software that are used to analyze and simulate phenomena in one field can be utilized broadly. This will only take place if all share standards and underlying technical infrastructures. Although many of the mechanisms to support the best scientific computing are becoming available through commercial channels, there continue to be special needs that the commercial sector is unlikely to meet directly because of the market size and technological risks.

Scientists must have easy access to the finest tools from the commercial and advanced research sectors, without dampening their creativity and ardor to do even better. Individual researchers expend too much effort, frequently with insufficient knowledgeable computing assistance, to create and re-create computing resources; to access, reformat, and save information; to protect the data and software assets. Much of this work could be done by computing experts and shared across the scientific research community. The ACP will encourage groups of scientists to undertake large coordinated information-intensive projects that can radically change the way they and their peers work, and that will support the sharing and long-term use of information that results from their work.

In summary then, the opportunity is here to create cyberinfrastructure that enables more ubiquitous, comprehensive knowledge environments that become functionally complete for specific research communities in terms of people, data, information, tools, and instruments and

that include unprecedented capacity for computational, storage, and communication. Such environments enable teams to share and collaborate over time and over geographic, organizational, and disciplinary distance. They enable individuals working alone to have access to more and better information and facilities for discovery and learning. They can serve individuals, teams and organizations in ways that revolutionize *what they can do*, *how they do it*, and *who participates*.

Figure 2.1 illustrates the types of facilities and services to be provided in an integrated way by a cyberinfrastructure layer (shaded). This layer is built upon base technology for computation, storage, and communication. Cyberinfrastructure should be produced and managed in a way that enables research communities/projects to tailor efficient and effective application-specific, *but interoperable*, knowledge environments for research and education. Interoperability is important for facilitating multidisciplinary projects as the evolution of discovery dictates. The Panel has learned that new types of scientific organizations and supporting environments ("*laboratories without walls"*) are essential to the aspirations of growing numbers of research communities/projects and that thus they have begun creating such environments under various names including *collaboratory, co-laboratory, grid community, e-science community,* and *virtual community*. The NSF through an ACP can now enable, encourage, and accelerate this nascent grass-roots revolution in ways that maximize common benefits, minimize redundant and ineffective investments, and avoid increasing barriers to interdisciplinary research.

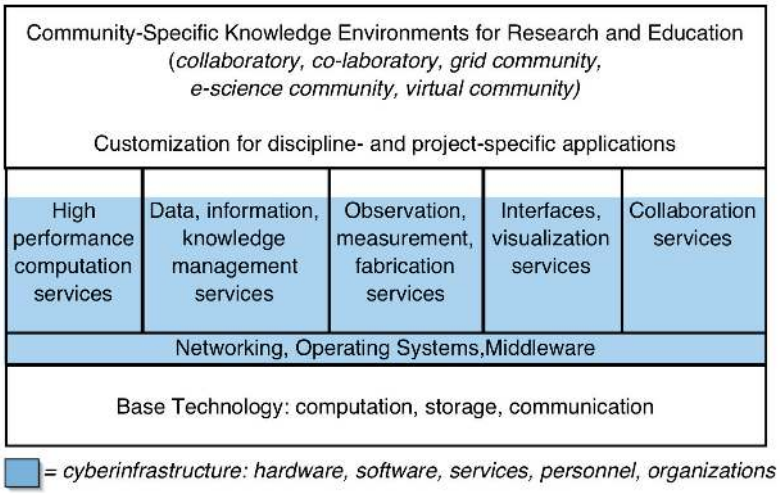

Figure 2.1 Integrated cyberinfrastructure services to enable new knowledge environments for research and education.

Achieving this vision challenges our fundamental understanding of computer and information science and engineering as well as parts of social science, and it will motivate and drive basic research in these areas. We envision radical improvements in cyberinfrastructure and its impact on all science and engineering over time, as work ripens at the intersection of fundamental technical and social *research* relevant to cyberinfrastructure, as well as the *application* of cyberinfrastructure to discovery and learning. Success in this venture has profound broad implications for research, education, commerce, and the social good.

2.3 Improving Information Technology Performance and Use

The vision of an ACP cannot be achieved by procuring existing commercial technologies alone. Of course, to the extent that commercial technologies and services are available off the shelf, they should be incorporated. But information technology is hardly mature; in fact, it is always evolving toward greater capabilities. Its applications are even less mature, and there are many opportunities to mold it to better meet the needs of end users. While possessing many commonalities with commercial technologies and applications in widespread use, science and engineering research have distinctive needs. These needs can often serve as technology drivers requiring extremes of processing and communication rates, storage capacities, the need for unanticipated access to data by many, and the longevity of data. Thus, research in new information technologies and applications utilizing those technologies often have important commercial spin-offs. This situation is illustrated by supercomputing, first applied to scientific and military applications and later to many commercial purposes.

The NSF mission includes advancing information technologies and their effective application to societal needs through basic and applied research in information technology. The ACP offers a significant opportunity for research into the more effective applications of information technology and opportunities for identifying and refining its supporting cyberinfrastructure. Just as supercomputing and numerical methods have been greatly advanced (and will continue to be advanced) by addressing the needs of the scientific and engineering communities, the ACP will be a significant driver for a diverse suite of technologies including collaborative technologies, massive interoperable distributed databases, digital libraries, and the preservation and mining of data. We expect (and the NSF should encourage) commercial spin-offs from this research, benefiting [21]commercial science and engineering research and development and other application areas.

The conduct of science and engineering is a social activity, pursued by individuals, collaborations, and formal organizations. Any enlightened application of information technology must take into

account not only the mission of science and engineering research but also the organizations and processes adopted in seeking these missions. A major opportunity in the ACP is to rethink and redesign these organizations and processes to make best use of information technology. In fact, this is more than an opportunity; it is a requisite for success. Experience has shown that simply automating existing methodologies and practices is not the most effective use of technology; it is necessary to fundamentally rethink how research is conducted in light of new technological capabilities. Advanced cyberinfrastructure offers the potential to conduct new types of research in new ways. Doing this effectively requires holistic attention to mission, organization, processes, and technology. It creates the need to involve social scientists as well as natural scientists and technologists in a joint quest for better ways to conduct research.

2.4 Rationale for Government Investment

The ACP requires government investment in research and development of cyberinfrastructure technologies (principally software) for several reasons. First, the marketplace under invests in long time-horizon research. The cyberinfrastructure and application technologies are within the domain of NSF responsibility for government-funded research, and the ACP will maintain U.S. leadership in these technologies through research, experimentation, and commercialization. Second, infrastructure and applications suffer from a chicken-and-egg conundrum that infrastructure requires a diversity of successful applications for its commercial viability, while commercial applications target only widely deployed infrastructure. This ACP will follow the successful model of the Internet, with targeted and coordinated government investment in infrastructure and applications, experimentation and refinement in actual uses, and coordinated commercialization of both elements together. Third, while we expect many if not most of the technologies developed in this ACP to be of broad applicability, science and engineering research has special needs in functionality, performance, and scale that are unlikely to be fully served by commercial firms, at least not without government assistance.

2.5 Scope of the ACP

We propose a large and concerted new effort, not just a linear extension of the current investment level and resources. NSF must recognize that both the scope and the scale of shared cyberinfrastructure must be far broader and deeper than in the past. Cyberinfrastructure includes computing cycles, but also broadband networking, massive storage, and managed information. Even these

are not sufficient. There must be leadership on shared standards, middleware, and basic applications for scientific computation. The individual disciplines must take the lead on defining certain specialized software and hardware configurations, but in a context that encourages them to give back results for the general good of the research enterprise and that facilitates innovative cross-disciplinary activities in the near term and in the distant future.

A major point is that cyberinfrastructure includes more than high - performance computing and connectivity. Not only is it focused on sharing and efficiency and making greater capabilities available across the science and engineering research communities, but it also serves other important goals such as facilitating new applications, allowing applications to interoperate across institutions and disciplines, ensuring that data and software acquired at great expense are preserved for future generations and easily available to all, and empowering enhanced collaboration over distance and across disciplines.

To succeed, NSF must institute a broad and deep program that supports the true needs of all the science and engineering missions within NSF by committing to make the fruits of cyberinfrastructure research and development (as well as related work from other agencies and companies) available in an integrated fashion to facilitate new approaches to scientific and engineering research. It must ensure that the exponentially growing data is collected, curated, managed, and archived for long-term access by scientists (and their IT applications) everywhere, to create and continually renovate a new “high end”, so that selected research projects can use centralized resources 100-1000 times faster and bigger than are available locally. The continuing geometrical improvements in computing speeds and storage and networking capacity mean that research groups and universities now have immediate access to far more resources than ever, but the recent limited national investment in high-end resources constrains the most aggressive research projects from achieving the next level of complexity and resolution.

National needs for advanced cyberinfrastructure will drive significant new research and development in computer and systems architecture. The NSF needs to take advantage of and participate in such efforts to continually improve research cyberinfrastructure; and to support research in areas of computing science that are likely to have largest impact. Science and engineering educators can also use the new infrastructure to educate the next generations of scientists using best techniques, spanning disciplinary boundaries, and democratizing participation. It can enhance international collaboration and resource sharing.

The ACP involves significant educational dimensions in terms of both needs and outcomes. The research community needs more broadly trained personnel with blended expertise in disciplinary science or engineering, mathematical and computational modeling, numerical methods, visualization, and sociotechnical understanding of grid or collaboratory organizations. Grid and collaboratory environments built on cyberinfrastructure can enable people to work routinely with colleagues at distant institutions, even ones that are not traditionally considered research universities, and with junior scientists and students as genuine peers, despite differences in age, experience, race, or physical limitations. These environments can contribute to science and engineering education by providing interesting resources, exciting experiences, and expert mentoring to students, faculty, and teachers anywhere there is access to the Web. The new tools, resources, extensions of human capability, and organizational structures emerging from these activities will eventually have beneficial effect on the future of education at all levels[24] and on knowledge-based institutions more generally.

The ACP also has great potential to empower people who, because of physical capabilities, location, or history, have been excluded from the frontiers of scientific and engineering research and education.

2.6 How Will Science and Engineering Research be Changed?

The vision of ACP is to use cyberinfrastructure to build more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools, and instruments and that operate at unprecedented levels of computational, storage, and data transfer capacity. Increasingly, new types of scientific organizations and supporting environments for science are essential, not optional, to the aspirations of research communities and to broadening participation in those communities. They can serve individuals, teams and organizations in ways that revolutionize *what they can do*, *how they do it*, and *who participates*.

Early computational models of physical, mechanical, and biological systems were confined to basic representations of the most fundamental properties and processes. Results from such models, based upon a limited number of calculations painstakingly evaluated, provided new theories and explanations of behaviors either observed in nature or simulated with physical models. Later, increases in computing and networking capabilities bred a new generation of models containing substantially greater realism and the ability to approach scientific and engineering problems from a “systems” point of view. Further, and perhaps more significantly, these models have led to fundamental discoveries.

The ACP is expected to produce another significant step forward in scientific and engineering discovery, not only through investments in raw computing, storage, and networking resources, but also by creating an infrastructure of equipment, software tools, and personnel – appropriately administered – to facilitate the solution of complex, coupled problems involving massive data collection, computation, and analysis. Cyberinfrastructure as we envision it includes not only high-performance computation services, but also integrated services for knowledge management, observation and measurement, visualization, interaction, and collaboration.

While it is impractical and unnecessary to make detailed projections of the impact of ACP on all science and engineering disciplines, a few examples can illustrate how scientific and engineering research will be revolutionized and the benefits that will flow from those changes.

Atmospheric Science – In 1998, the National Research Council[25] noted that, although small- and intermediate-scale climate modeling in the United States is enjoying notable success, the highest-end research opportunities are limited in part by the lack of appropriate computing resources. Not surprising, the highest-end resources are most important for understanding the carbon cycle and other complex processes that govern the global climate system. The ACP will enable the development and execution of fully coupled Earth system models that will allow the simulation of climate for hundreds and thousands of years, down to grid spacing of 10 km, and that will include complete and fully linked representations of chemical, biological, and ecological processes in the atmosphere, hydrosphere, and lithosphere.

At the other end of the time spectrum, today's operational global and hemispheric weather forecast models utilize grid spacing of ~50 km, while limited-area regional/synoptic models operate on grids of 15-20 km spacing. Although such representations of the atmosphere are vastly better than those used even a decade ago, they remain inadequate for capturing nature's most intense and locally disruptive weather. Research now under way in the explicit prediction of individual thunderstorms and their wintertime counterparts, using grid spacing of order 1 kilometer, is showing considerable promise and could have a tremendous impact on aviation, communications, agriculture, and energy. However, the computational challenges are daunting. The ACP will enable research to create effective frameworks for both exploring small-scale atmospheric predictability and dealing with their associated massive amounts of observational data and model output. It will also enable the federation of the necessary multidisciplinary, multi-institutional, and geographically dispersed human expertise, archival data, and computational models.

Forestry – Tremendous progress is being made in the modeling of wildfires, with the explicit inclusion of fuels and chemical reactions and full two-way coupling with the atmosphere. The ACP computational

resources will allow for a more complete representation of land-surface characteristics, fuel composition and consumption, and feedbacks, leading to more effective strategies for combating fires including the development of chemical agents whose impacts can be tested empirically, at very large scales, in a virtual world. We can also imagine the emergence of "rapid response collaboratories" that will eventually enable an actual forest fire to be modeled in real time based upon sensor data from the field and used to monitor and direct the process of fire fighting. Running in a predictive mode, these models demonstrate the capability to anticipate a "blowout" event in time to move a fire fighting crew to safety.

Ocean Science – The field of ocean sciences is poised to capitalize upon extraordinary opportunities for advancement, ranging from an understanding of the roles played by the world's oceans in climate and global change to the delicate balances that exist in coastal ecosystems. Computer ocean models now are capable of simulating detailed turbulent structures and transport processes in three dimensions. To understand and predict the full climate system will, for example, require facilities in the ACP capable of computational coupling to atmospheric models, and inclusion of the complex chemistry needed to understand the physics of carbon sequestration. Such efforts require computational, data, and networking resources orders of magnitude beyond those presently available. The coastal zone, fundamentally important to fisheries, defense, recreation, and human health, is a vastly complex environment affected by freshwater runoff, the introduction of large inputs of nitrogen and other nutrients, and the episodic release of pollutants. An accurate representation of these and other biogeochemical processes will allow for better stewardship of the coastal environment and provide frameworks for policy decisions affecting the nation's economy.

Environmental Science and Engineering – The previous three activities and many others are part of a growing collection of interdisciplinary and interorganizational activities in the area of environmental research and education, much of it nurtured by a cross-cutting Environmental Research and Education (ERE)[26] program at the NSF. This community has been among the leaders in exploring requirements for cyberinfrastructure supporting the necessary integration of environmental research and education focused on understanding fundamental processes involved in physical, biological, and human system interactions. Examples include research in the areas of ecosystem dynamics, cell function, atmospheric chemistry, biogeochemical cycles, political or economic institutional processes, coastal ocean processes, population biology and physiological ecology, Earth system history, solar influences, and the study of the interactions responsible for the ozone hole. This is an example of a community for which advanced cyberinfrastructure will have a high payoff.

Space Weather – Although terrestrial weather and climate are of considerable importance to society, space weather – or the conditions in space that arise from interactions between the Earth and sun – is growing rapidly in importance. Active space weather can, for example, disrupt surface and space-based power and communication infrastructures, benefiting not only commerce and the economy but also national defense. To date, the sun and Earth have been studied largely as individual, isolated systems. However, a fully coupled Earth-sun framework is essential for understanding the physics and societal impacts of space weather. This is a truly global research community, and the ACP will create a collaboratory of international science teams, hundreds of ground and space-borne instruments, predictive computational models, and historical data archives. This will improve fundamental understanding and operational space weather forecasting.

Computer Science and Engineering – The foundation of cyberinfrastructure is computer and information science and engineering – areas whose breadth and impact have expanded in the past two decades, and upon which numerous other disciplines depend for efficient and reliable processing, communication, security, management, storage, and visualization. The research challenges are varied, and enable revolutionary science and engineering. Unconventional architectures based upon new substrates (e.g., quantum and biological, including smart fabric and molectronics) offer promise for breaking the silicon CMOS barrier. Self-diagnosing and adaptive systems will be essential for managing the increasingly complex distributed hardware and software infrastructures. We also face challenges in security, scalability, fault tolerance, brokering, scheduling, and policy. Digital libraries, metadata standards, digital classification, and data mining are critical. Additionally, more effective languages, compilers, middleware, and integration – especially in utilizing distributed systems – are a key enabler. An ACP could revolutionize computer science and engineering research itself because, for example, of its inherent complexity and requirements for systemic integration, the opportunity for synergy between creating and applying new knowledge, and the need for a more integrated understanding of the technical and social dimensions of cyberinfrastructure applied to research and education.

Information Science and Digital Libraries – An information-driven digital society requires the collection, storage, organization, sharing, and synthesis of huge volumes of widely disparate information and the digitization of analog sensor data and information about physical objects. The digital library encompasses these functions, and research and development are needed for the infrastructures to mass-manipulate such information on global networks. Digital libraries also provide powerful tools for linking and relating different types of information, leading to new knowledge. These capabilities require new paradigms for information classification, representation (e.g., standards, protocols, formats, languages), manipulation, and visualization. The ACP will spearhead such new developments.

Biology/Bioinformatics – A new era of biology is dawning exemplified by the human genome project and the promise of new science affecting areas such as crop production and personalized medicine. The raw DNA sequence information deposited in public databases doubles every six months or so; its analysis has motivated development of the new field bioinformatics. Characterization of protein folding, for example, utilizing the 30,000 or so protein structures currently available in the public repository, would require hundreds of years on today's desktop computers. While this calculation can be completed in weeks on currently available massively parallel teraflops computers, under the envisioned ACP one can imagine the process being reduced to literally hours. Such work will improve our understanding of myriad biological functions and disease states and provide a framework for developing new therapies and disease and weather/climate resistant plants.

Medicine – Medical advances of great benefit to humanity are expected, ranging from telemedicine and drug therapies to non-invasive repair of damaged tissue. A significant breakthrough will be the creation of a functional, three-dimensional cyber human body. This capability will provide vast educational opportunities ranging from the performance of surgeries on virtual cadavers to physiology education of middle- and high-school students. Much like flight simulators, the virtual human also provides a framework for repeated experimentation under strictly controlled conditions, ranging from macroscale structures (like organs and the musculo-skeletal system) to individual cells. Among other benefits, a virtual human will significantly reduce the imbalance among schools in their facilities for studying human physiology.

Physics – Physics is pursuing major projects depending on advanced cyberinfrastructure. High-energy physics, for example, must have global-scale, high-performance grids and collaboratories to support the acquisition, distribution, storage, and collaborative evaluation of the massive data sets generated by the premiere instruments at CERN. Global scale collaboration will enable experimentation and also designing and constructing facilities and the experiments using them. This community is using cyberinfrastructure to support distributed learning for professional development and to allow faculty to remain active in teaching and mentoring at their home institutions while resident at CERN in Switzerland.

Astronomy – Traditionally, astronomers have analyzed observations of individual targets while assembling theories limited in their consideration of larger-scale interactions. Such individual observations are being replaced by whole-sky surveys of enormous detail and petabyte datasets, providing global views of phenomena ranging from black holes to supernovae, and identifying new objects so rare that only one or two may exist among billions of objects. The needed computational infrastructure does not exist but will be enabled by the ACP. This revolution in astronomy driven by cyberinfrastructure

promises to enable a whole new level of understanding of the universe, its constituents, and their origins and evolution, touching on the issues ranging from the fundamental physics in the early universe to the abundance of Earth-like planets and the origins of life.

Engineering – The distinctions between science and engineering are blurring, as illustrated by an engineering component in all the areas in this section. One example of the impact on engineering practice is the understanding of turbulence. Thirty years ago, it was generally believed impossible to perform direct numerical simulation of turbulence (i.e., simulation from first principles, with explicit representation of chaotic motions). Today this has been done, and is revolutionizing the design of combustion engines, aircraft, and automobiles as well as the understanding of clouds and the spread of pollution. However, it is not currently possible to simulate turbulence in large volumes or at high speeds – a significant limitation affecting most of the interesting and relevant applications. Further, the massive datasets produced by direct turbulence simulations are difficult to analyze and visualize, thus thwarting efforts to move from the turbulence produced by a small bird to that produced in the wake of a jumbo jet. The ACP will make available the raw computational, data handling, and visualization resources needed to meet these challenges and thus to improve the manufacturing of large and small devices where turbulence is important.

Materials Science & Engineering – Computer simulations, enabled by the envisioned ACP, will make possible quantum mechanical calculations on nanoscale systems, which, in turn, will enable the fundamental principles governing the rational design of new materials for nanotechnology to be uncovered. Such simulations will, for example, contribute not only to the design but also to the rational synthesis of truly novel materials for IT and national security applications and of nanocatalytic materials for the chemical industry. Extrapolation of what is currently possible with simulations based on classical mechanics and atom-based force fields on current teraflops computers indicates that structural and dynamical properties of trillion-atom systems covering a length scale of a few microns will be possible on a petaflops computer. This will enable the study of systems ranging from nanoscale composite materials with realistic microstructures to biologically inspired self-assembled devices for medical applications. Further, multiscale quantum-atomistic-continuum simulations using the envisioned ACP may enable the integration of thousands of heterogeneous teraflops-scale physical models that will be needed for more fundamental component design and optimization in advanced engineering applications.

Social and Behavioral Sciences – As a relatively new user of cyberinfrastructure, the social and behavioral sciences are poised to make tremendous advances in a variety of areas ranging from cognition and linguistics to economic forecasting. Simulations of the

interplay between concepts and perception in the course of analogy making have been created, and programs are under development for modeling the perception and creation of style in the world of letterforms. Devices that convert neural signals to speech are being studied, and new techniques based upon numerical simulation are accelerating the pace of mental and physical rehabilitation, particularly for cases of extreme physical trauma. New virtual organizations and practices made possible by cyberinfrastructure provide new areas of study within the social sciences.

2.7 Participation Beyond the NSF Community

NSF has both a unique breadth of scientific scope and responsibility for the health of the scientific research enterprise in the U.S., so NSF is ideally poised as a leader in cyberinfrastructure within the federal government. However, ACP cannot be fully effective if it is an NSF-only program: significant coordination with other federal agencies, universities, industry, and international programs is required. This will magnify the impact through interoperability and consistency across a larger universe of researchers and will also bring significant added resources to bear.

Other Research Sponsors – The NIH is spending billions of dollars annually on information technology infrastructure and its support and use in research, but in a way that may not lead to a common, interoperable cyberinfrastructure, nor infrastructure at the leading edge. NIH has recently initiated more coordination, in the spirit of an ACP, for example the Biomedical Informatics Research Network (BIRN)[15]. Similarly the Department of Energy (DOE) National Collaboratories Program[16], and the DOE program in Scientific Discovery through Advanced Computing (SciDAC)[17] are examples of growing investment in cyberinfrastructure that can supplement NSF investments.

Industry and Universities – Some of the capabilities needed in ACP are commercially available, industry may be interested in developing new technologies relevant to the science and engineering research community, and many technologies that are an outgrowth of the ACP research and development will be of interest to the commercial sector. Thus industry must be a partner in development and deployment in the ACP and will also be a beneficiary. ACP will also encourage co-investment by universities in advanced cyberinfrastructure on campuses and will provide models and experience with new tools and new organizational forms for knowledge creation and education in the digital age. It could directly complement, for example, a major three-year study now begun at the U.S. National Academies of Science, Engineering, and Medicine on information technology and the future of the research university[24]. It can catalyze and provide over-the-horizon visibility to other agencies, research labs, and education-at-large.

International – It is imperative that the ACP interoperate with cyberinfrastructure being developed and deployed in other countries. Science is international; and many other countries have expertise, data resources, computing systems, applications and systems software, and instruments (such as telescopes and particle accelerators) that need to be available more easily to international teams including American scientists (and vice versa). The high-energy physics community, for example, must have appropriate cyberinfrastructure to enable collaboration in experiments using the premier instruments at CERN in Switzerland.

Collaboration within and among disciplines is growing rapidly; in some cases hundreds of scientists are working on a single project across the globe. Cyberinfrastructure must support this type of collaboration in a reliable, flexible, and cost-effective manner. The activity in Europe and Asia in cyberinfrastructure has increased of late; it is mutually beneficial to established strong links with relevant international efforts and to co-fund significant collaborative international projects. Science is increasingly global, yet it is still difficult to fund joint international e-science projects that develop or require cyberinfrastructure.

Major scientific laboratories elsewhere have contributed significantly to advanced scientific computing, and continue to do so. (The Web was born at CERN, just as the browser was born at NCSA.) For example, the UK National Grid is part of their overall e-science effort, and the Netherlands National Grid has similar goals. The EU is considering a number of even broader Grid proposals.

A few examples of relevant international activities include the following:

- The UK recently launched an "e-Science" program[18] that has many of the characteristics of the ACP. The aims of this program include:

  – provide infrastructure and facilities needed for next major stages of international collaborative research in genomics and bioscience, particle physics, astronomy, earth science & climatology, engineering systems, and the social sciences;

  – contribute to the emergence of next generation open platform standards for global information utilities;

  – solve major challenges in processing, communication, and storage of very large volumes of valuable data;

  – provide *generic* solutions to needs of individual disciplines and applications; and

  – provide optimal international infrastructure.

Initial funding for the e-Science program is on the order of $200 million over three years, most of which is allocated to large applications projects and a quarter of which is devoted to developing the necessary software infrastructure. The latter efforts are collaborating closely with American and European projects that are developing middleware and in some cases even providing funding for those international groups. The e-Science funds are supplemented by infrastructure funding from previously existing programs that support both a very capable UK-wide research network (10 Gb/s backbone) and high-speed international links and high end computing resources. On the latter topic, in July the UK Science Research Council signed a contract with IBM with an overall cost of £53m (over $82M) for a computer system known as HPC(X). The initial 3 teraflops configuration of HPC(X) will ramp up to 12 teraflops by 2006, with a teraflops rating based on LINPACK performance.

- The European Union has funded well over a dozen Grid projects as well as a high-speed European research network – GEANT[27]. GEANT reaches over 3,000 research and education institutions in 30-plus countries through 28 national and regional research and education networks. It is also quite fast: nine of its circuits operate at speeds of 10 Gbps, while eleven others run at 2.5 Gbps. GEANT has the dual roles of providing an infrastructure to support research in application domains and providing an infrastructure for (network) research itself.

- In the upcoming Sixth Framework Program[19], the EU has allocated 300M euros for further upgrading the GEANT network and for building large-scale Grid test-beds. A solicitation for proposals will be issued in the first half of 2003. In addition, there are a number of grid projects under way funded by individual countries. A partial list includes Canada, China, Denmark, India, Japan, Korea, Norway, Romania, Sweden, and Switzerland. Typical funding levels are tens of millions of dollars per project over several years.

- Other countries also have significant computing resources that are used for computational science. In the early 1980s U.S. academic researchers gained access to European computing facilities enabling larger-scale computational science research. In Japan, many universities and research laboratories have high-end facilities, and in March 2002 the Earth Simulator[20] system became operational. The Earth Simulator, currently the world's fastest computer system with a peak speed of 40 teraflops, was built by NEC for the Earth Simulator Research and Development Center, a collaborative organization of the National Space Development Agency of Japan, Japan Atomic Energy Research Institute, and Japan Marine Science and Technology Center. The Earth Simulator is targeted at analysis of global environmental problems through simulation of geophysical, climate, and weather-related phenomena.

At present 55% of the top 500 computer systems in the world (based on LINPACK ratings), representing 56% of the aggregate LINPACK flops, are outside the US. Of the top 100 systems, 33 are designated for academic use. Of those, only 9 are in the US, even if one includes the systems at NCAR[28] and at NERSC[29]. In areas such as high-end computing and high-speed network infrastructure, other countries are either in the lead or on a par with the US. However, late in 2002 Lawrence Livermore National Laboratory announced an order to IBM for delivery of a 100 teraflops machine in 2004 (for national security calculations) and delivery of a 360 teraflops machine in 2005 (mostly for open scientific applications). Many scientific investigations have international components and therefore ACP should make both U.S. and international resources available for shared international collaboration.

2.8 Educational Needs and Impact

A new interdisciplinary work force – The need for a new workforce – a new flavor of mixed science and technology professional – is emerging. These individuals have expertise in a particular domain science area, as well as considerable expertise in computer science and mathematics. Also needed in this interdisciplinary mix are professionals who are trained to understand and address the human factors dimensions of working across disciplines, cultures, and institutions using technology-mediated collaborative tools. Prior work on computer-supported collaborative work and social dimensions of collaboratories needs to be better codified, disseminated, and applied in the design and refinement of new knowledge environments for science based on cyberinfrastructure.

The term “computational science and engineering” (CSE) has emerged as a descriptor of broad multidisciplinary study that encompasses applications in science/engineering, applied mathematics, numerical analysis, and computer science. As noted by the Society for Industrial and Applied Mathematics (SIAM)[30]:

Computer models and computer simulations have become an important part of the research repertoire, supplementing (and in some cases replacing) experimentation. Going from application area to computational results requires domain expertise, mathematical modeling, numerical analysis, algorithm development, software implementation, program execution, analysis, validation and visualization of results. CSE involves all of this.

SIAM notes that “CSE is a legitimate and important academic enterprise even if it has yet to be formally recognized as such at some institutions. Although it includes elements from computer science, applied mathematics, engineering and science, CSE focuses on the integration of knowledge and methodologies from all of these disciplines, and as such is a subject which is distinct from any of them.”

The community surveyed in creating this report noted repeatedly that insufficient attention is being given to educating non-computer or domain science students in the concepts and tools of cyberinfrastructure. For example, graduate and higher-level undergraduate courses in computer science are designed for disciplinary majors, and non-majors wishing to take such courses are dissuaded by an onerous prerequisite structure. Further, even if such skills are attained, domain science courses often do not exercise them sufficiently, leading to atrophy of skills.

In response to this problem – while also recognizing the need to maintain strong, traditional disciplinary programs in science and engineering research – significant resources must be directed toward developing programs of study in the computational sciences at both the graduate and undergraduate levels. A survey of educational objectives, as well as sample programs and curricula, can be found at the SIAM Web site.[30]

Continuing education is also needed. Community-wide workshops are needed for science and engineering practitioners so as to function effectively in the rapidly evolving IT world. Such workshops and courses could be delivered via distance learning and would lower the entry threshold for those new to high-performance computation.

Impact on science and engineering education – The ACP requires the aforementioned innovation and reforms in education and can also be directly leveraged in science and engineering education. Grid and collaboratory environments built on cyberinfrastructure can enable people to work routinely with colleagues at distant institutions, even ones that are not traditionally considered research universities, and with junior scientists and students as genuine peers, despite differences in age, experience, race, or physical ability. These new environments can contribute to science and engineering education by providing interesting resources, exciting experiences, and expert mentoring to students, faculty, and teachers anywhere. By making access to reports, raw data, and instruments much easier, a far wider audience can be served. Since broadband networks are increasingly available in schools, videos and other complex effects can be viewed by students and teachers as well as by researchers.The new tools, resources, human capacity building, and organizational structures emerging from these activities will also eventually have even broader beneficial impact on the future of education at all levels, in almost all disciplines, and in all types of educational institutions.

Minority Serving Institutions – An important goal of the ACP must be to more effectively include Minority Serving Institutions (MSIs), which include Historically Black Colleges and Universities (HBCU), American Indian Tribal Colleges (AIT), and Hispanic Colleges and Universities (HCUs) and other underrepresented groups into mainstream scientific and engineering research and education. Few of these institutions were involved in discussions leading to the original NSF supercomputing centers, and collaboration efforts to date, though well intentioned and covering a spectrum of activities ranging from education/outreach/ training to basic research, have for the most part fallen short of their goals for a variety of reasons. This failure is particularly troubling in light of the fact that, by 2035, it is estimated that one in five Americans will be Hispanic.

One of the most important barriers to engaging MSIs in research using cyberinfrastructure is the lack of adequate network connectivity – a problem especially acute for the Tribal Colleges because of their largely rural location and frequently impoverished localities (three of the five poorest counties in the United States are homes to Tribal Colleges). Further, such institutions lack the tools and infrastructure needed to participate in mainstream research. Although various initiatives (e.g., EOT-PACI[31], the Advanced Network with Minority Serving Institutions Initiative[32]) have shown promise, the principal audience has been IT staff rather than faculty and researchers. These and other limitations have perpetuated the so-called digital divide, reflected by a 20+ year gap in capability between mainstream institutions and many MSIs (based on statistics from the U.S. Department of Commerce, 46.1% of white non-Hispanic households have access to the Internet, compared with 23.6% for Hispanics).

Although this challenge is multifaceted, solutions need not be incrementally applied; indeed, it is eminently possible, through a significant infusion of both technology and education, to close the digital divide and establish meaningful research collaborations and educational initiatives. The PITAC[33] emphasized the importance of reaching MSIs, and we underscore it again here. The ACP therefore must support strategic IT planning for underserved communities. In addition, opportunities for research collaboration must be more effectively communicated to both mainstream institutions and MSIs, and significant efforts must be directed toward engaging underserved communities directly, rather than as programmatic add-ons.

Experimental Program to Stimulate Competitive Research – A more encouraging story can be told about EPSCoR[34] (Experimental Program to Stimulate Competitive Research), which at present involves 21 states and the Commonwealth of Puerto Rico. A joint program of the NSF and several U.S. states and territories, EPSCoR promotes the

development of science and technology resources through partnerships involving universities, industry, government, and the federal research and development enterprise. It operates on the principle that aiding researchers and institutions in securing federal R&D funding will develop a state's research infrastructure and advance economic growth, and its main goal is to maximize the potential inherent in a state's science and technology resources and use those resources as a foundation for economic growth.

Most EPSCoR states have taken significant steps to provide high-speed connectivity and engage researchers in collaborative cyber activities with major universities and national centers and laboratories, and these efforts should be continued and expanded. For example, the University of Kentucky spearheaded a project through which scientists and researchers in EPSCoR states can use Access Grid (AG) technology to bridge the digital divide caused by their geographic dispersion and limited funding. Six EPSCoR-grant states are implementing AG nodes, and the two newest EPSCoR states, Hawaii and New Mexico, have nodes as a result of their participation in the National Computational Science Alliance.

EPSCoR co-funding of mainline research grants, particularly in the Information Technology Research Program, has had a significant positive impact on research competitiveness of participating institutions (see http://www.ehr.nsf.gov/epscor/start.cfm). In the NSF Geosciences Directorate alone, EPSCoR co-funding has increased by a factor of 3 in the past few years. The ACP should embrace EPSCoR and continue to support what clearly is a very successful, high-impact program. Indeed, the EPSCoR model could be applied more specifically to MSIs, particularly with regard to high performance network connectivity.

Access by the wider public – By making access to reports, raw data, and instruments much easier, a far wider audience can be served. Although large teams and major financial investment are required to create comprehensive data repositories and specialized scientific facilities, individuals, even amateurs, working alone or in small groups, given access to such resources, can provide scientific discoveries. A good example is amateur astronomy, which significantly expands the reach of scientific observation.

Participation by the physically challenged – There are many ways to assist scientists and other users who have physical constraints through advanced cyberinfrastructure, as long as this opportunity is addressed from the beginning. Most of these resources are likely to be provided close to the individual rather than in a shared environment. Many of these supportive pieces of hardware and software will be generic, but there may be some tools specific to the scientific milieu. We have also identified a few functions that could most appropriately be implemented centrally. A few examples that should be considered for implementation within the ACP will illustrate this. Even people who are not challenged

will still find some of these features of use – a common observation about assistive technologies, the so-called curb-cut effect (sidewalks with curb cuts are simply better sidewalks – they help bikers, skaters, and people pushing strollers – not just those confined to wheel chairs).

It takes massive computing to do a first-class conversion of speech to text. (Online and moderate accuracy conversion can be done by commercial software on a typical PC, but higher accuracy requires elaborate algorithms that repeatedly examine delayed inputs.) Such computing might be provided as a Grid service on shared multiprocessors and would make an excellent adjunct for collaborative environments such as the Access Grid[35]. By using a shared networked resource, the service would be available to hearing-impaired scientists wherever they are. (The service would also be valued for making seminars available for delayed use by everybody.)

Infrastructure services that can convert sounds to visual signals would help the hearing-impaired interact with experimental equipment. The inverse translation of control panels to sounds would be useful for the visually impaired. This specialized translation does not fit simply into the commercial Webpage-enablement paradigm and may be particularly important for control gauges and warning devices. There could be broad social benefit to providing standards and support software for infrastructure-connected apparatus. (Sighted people might benefit from audible alarms, and workers in crowded environments might prefer silent visual signals, so standardized conversions for laboratory equipment may find broader usage.)

A research challenge would be to extend "visualization" to provide information for the visually impaired. Tactile (haptic) exploration combined with audible signals may be a useful way to convey information about complex phenomena and mathematical surfaces. If successful approaches are found, they should be made available through the ACP.

Digital libraries, discipline-specific collections, and archives of the published literature will be key components of the cyberinfrastructure. It is difficult for people with motor or visual disabilities to point to many specific items, or to look at very long stretches of text. A variety of services would help them and would also speed the work of others. Some possible approaches are abstracting services and interfaces that encourage skipping or shifting focus.

2.10 Overall Finding and Recommendation

The Panel's overarching finding is that a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology; and pulled by the expanding complexity, scope, and scale of today's research challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive "cyberinfrastructure" on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy. The cost of not doing this is high, both in opportunities lost and through increasing fragmentation and balkanization of the research communities.

Such environments and organizations, enabled by cyberinfrastructure, are increasingly required to address national and global priorities such as understanding global climate change, protecting our natural environment, applying genomics-proteomics to human health, maintaining national security, mastering the world of nanotechnology, and predicting and protecting against natural and human disasters, as well as to address some of our most fundamental intellectual questions such as the early formation of the universe and the fundamental character of matter.

As will be discussed in Section 5, there is already a significant base of effort and capability in the PACIs, which were created in response to the Hayes Report[36]. They run computing and data centers, create important middleware and scientific software, and coordinate activities with other scientists. Subject to appropriate review, we anticipate that they will play a continuing but evolving substantial role in the greatly enlarged activity we propose.

The Panel's overarching recommendation is that the National Science Foundation should establish and lead a large-scale, interagency, and internationally coordinated Advanced Cyberinfrastructure Program (ACP) to create, deploy, and apply cyberinfrastructure in ways that radically empower all scientific and engineering research and allied education. We estimate (details in Section 6) that sustained new NSF funding of $1billion per year is required to achieve critical mass and to leverage the necessary coordinated co-investment from other federal agencies, universities, industry, and international sources required to empower a revolution.

This Panel believes that the National Science Foundation has a once-in-a-generation opportunity to lead the revolution in science and engineering through coordinated development and expansive use of cyberinfrastructure.

The following sections of this report provide a further basis for this recommendation, our estimate of new funding required, and principles for the organization and management of the program. Appendixes provide additional details.

3.0 Trends and Issues

In this section we describe the converging and reinforcing trends and issues derived from our surveys, testimony sessions, readings, and deliberations; these formed the basis for the vision, findings, and recommendations presented in the preceding section. As mentioned earlier and illustrated in Figure 3.1, the ACP opportunity derives from a combination of the push of technology trends and the pull of vision and needs for its application in research communities. The impact of these trends is not necessarily linear. As certain thresholds of functionality or price-performance are crossed, disruptive changes occur. The trends may also reinforce one another, magnifying their impact.

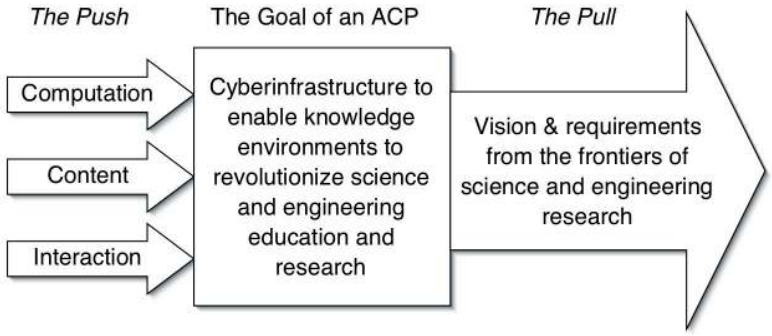

Figure 3.1. The push and pull for an ACP.

We have clustered these trends and issues into three areas: *computation*, *content*, and *interaction*. The substrate for all of these trends is the familiar exponential increase in the capacity of the base computation, storage, and communication technologies.

3.1 Computation

The measures of computing, networking and storage capacity continue to grow geometrically. We take for granted that computer speeds will rise radically with each new hardware generation, that machines will have more memory than before, that disks will hold ever more information, that the network will be faster, and that partially as a result software will provide ever more complexity and features. We should not hit physical limits for current basic chip and disk technologies before 2010 (and probably much later), so we assume continuation of this

golden age of information technology through the period addressed by this report. (Consideration of other technologies such as quantum computing is beyond the scope of this report, but research underway suggests that technology may move onto even higher performance curves in the future.)

As we have ridden these smooth exponential curves for several decades, what has changed? We have passed several practical thresholds, resulting in qualitative breakthroughs. Scientific research that would have been prohibitively expensive or previously demanded national-scale resources can be done in local facilities. Workstations can now do computations that only the biggest and most expensive supercomputers could attack a few machine generations ago. Thus, serious computations demanding real-time visualization, simulation of interactions of thousands of particles, and 2D- and even 3D fluid dynamics are possible on the desktop. Combining commodity hardware (PC boards and networks) into a laboratory cluster permits computations that only national labs could attempt a decade ago. The entire scientific literature can fit on a few hundred disks, with material costs under $25K. (Disk storage became cheaper than paper years ago and is also competitive with microfilm.) There are individual civilian laboratories and state universities that are installing computers in the teraflops range and data server clusters in the 100 terabyte range.

But the demand for highest-performance computation is also increasing, and thus we continue to need a hierarchy (or "pyramid") of connected computation resources of varying capacity and cost. In a few more years, we will cross the "peta" (10^{15}) line: there will be some supercomputers in the 0.1-1 petaflops range, some scientific databases will exceed 1 petabyte, and networks will exceed 1 petabits/s.

Hardware components – The hardware components underlying computation, storage, and communication have been improving exponentially (at a compound rate of growth) for many years; this is expected to continue over the scope of this report. Although the rate of growth of circuit speed may begin to flatten, major research directions have potential to break these barriers. The current speed growth directions are depending less on circuit speedup, and more on increasing circuit density and the number of parallel processing units on a chip or wafer. We will achieve petaflops not with a femtosecond clock, but rather by having a million processors on a nanosecond clock (give or take a factor of 10).

Processing – Computer speed is usually expressed as arithmetic calculations, or floating-point operations, per second (flops). In 1999, two machines in the world had a theoretical capacity of 1 teraflops. By now we estimate a dozen universities and laboratories have or have ordered computing clusters with theoretical capacities exceeding 1 teraflops, and by 2005 machines up to 10 teraflops will be relatively commonplace (a teraflops machine may even be affordable for

some individual researchers). These changes are due to continued improvements in chip technology and the ability to utilize clusters of chips and mass-produced computers. We benefit from not only parallelism, but also speed; in late 2002, a clock rate of 350 GHz was announced for a silicon-based experimental device.

Storage – Many applications depend on manipulating masses of data, far more than can reside inside the processors. These data can be observational inputs, experimental values, or results of calculations, images, or videos. Such information is usually kept on disk (though the largest archives are stored on removable optical disks or magnetic tapes). The highest performance (measured variously as total number of characters of information stored, number of characters per volume of lab space, or number of characters retrieved per second) is generally found in the most recent commercial disks. Increasing overall storage capacity comes from utilizing many disks to store massive amounts of information and accessing them in parallel.

Disk capacities (measured as bits per square inch of magnetic material) have historically increased at 60% per year, but in the past few years bit storage density has increased by about 100% per year. Prices of individual units have fallen more slowly, so most of the economic improvement has come from larger capacities. The most capacious disks in late 2002 store about $3x10^{12}$ bits (320 gigabytes, or 0.33 terabytes) of information. Databases of a few terabytes are common; only ones over 100 terabytes are considered remarkable.

Networks – A major shift in computing has come from the practical availability of high-bandwidth data networks. Network connections up to 45 megabits/s are easily available, connections over 155 megabits/s are still aggressive, and some research institutions are beginning to connect at 2.5 gigabits/s and faster. Available technology can support far higher bandwidths. Deployments have already demonstrated 1.6 terabits/s on a single fiber (40 channels at 40 gigabits/s), while laboratory experiments have reached over 11 terabits/s. Switching data at these speeds remains relatively expensive, but technologies have been demonstrated.

Network researchers and providers are also introducing a new paradigm –*optical networks* based on the emergence of an optical layer, operating entirely in the optical domain (and avoiding electronic bottlenecks), to enable very high capacity end-to-end wavelength ("lambda") services that provide (through wave division multiplexing) many virtual fibers on a single physical fiber. Optical networks, for example, are being explored for linking widely distributed high-performance machines together in grids. Optical networking is an important emerging technology to explore and use in the ACP.

These improvements make it plausible to move huge files between sites, so that computing and storage facilities can be split or combined in a number of ways. However, the speed of light (~1 ns/ft, or about 20 ms to cross the U.S.) is not increasing, and networking switches add further delays. This puts a fundamental limitation on the use of widely dispersed processing and storage resources for tightly coupled computations and is one of the reasons that supercomputers remain indispensable for many scientific applications.

In-building networks are improving in two ways – bandwidth and mobility. Local area network (LAN) technology is now moving to high-speed Ethernets able to deliver 100 megabits/s or 1000 megabits/s to the individual server or desktop. Few current computers can handle such data rates effectively, nor can typical laboratory switches manage many full-speed streams, but this situation will improve rapidly.

The use of wireless (radio) access to the network is exploding, both within buildings and in general public uses. Very local access (using for example the IEEE 802.11 family of standards) can provide many megabits per second to a single device (laptop or PDA), and new generations of cellular telephone technology will permit 0.1-1 megabits/s to the roaming device in the next half dozen years. This has great promise for many mobile applications, such as gathering scientific data in the field and geographic-independent group collaboration.

Displays –Typical commercial displays offer about 1 square foot of useful visual information and present around 1 megapel (million picture elements). This is another technology that is rapidly advancing. Many labs (especially those on the Access Grid) combine between 3 and 15 typical displays to present a single large image. Recent special displays have higher density and brightness; desktop devices with over 9 megapel are now commercially available. Displays are also configured to provide 3D and immersive virtual reality experiences in CAVES or ImmersaDesks. Costs continue to decline and very useful 3D interaction is now available below $10K.

Provision and use of high performance computing – World leadership in the highest-capacity computing has been, and continues to be, a significant factor in research and national security. The federal government has been the primary investor in, and user of, the highest capacity machines. Mission agencies such as the Department of Energy and the Department of Defense have used supercomputers in mission-specific domains, including some use in basic research. The NSF, however, is specifically charged with fostering and supporting broad development and use of computers and other scientific methods and technologies for broad research and education in the sciences and engineering.

As computers evolved, various NSF directorates supported research in components, theory, software, systems, and applications of computers. An Advanced Scientific Computing (ASC) Program, situated in the Office of the Director, provided the NSF research community access to the highest-performance supercomputers of the day. In 1985, ASC activities and several other programs were merged into the Directorate for Computer and Information Science and Engineering (CISE). CISE supports investigator-initiated research in all areas of computer and information science and engineering and also supports high-performance national computing and information infrastructure for research and education generally. It has done this through co-investment in computational infrastructure in academia, and at the high end, through a series of centers and alliances. The development and operation of high-performance computational centers was also instrumental in the creation of the NSFNET, the precursor of the commercial Internet. In addition, the recommendations of the 1995 *Hayes Report (Report of the Task Force on the Future of the NSF Supercomputer Centers Program)*[36] along with the predecessor *Branscomb Report*[37] *(NSF Blue Ribbon Panel on High Performance Computing)* formed the basis for the development of the Partnerships for Advanced Computational Infrastructure (PACI) program.

Two PACI[2] partnerships established in 1997 are currently operating under the principles set forth in the Hayes Report by (1) providing access to high-end computing, (2) affording knowledge transfer of enabling technology and applications research results into the practice of high-performance computing, and (3) supporting education, outreach and training activities. Each partnership consists of a leading-edge site, the National Center for Supercomputing Applications in Urbana-Champaign and the San Diego Supercomputer Center in San Diego, and a significant number of partners. The highest-capacity machines are located at the two centers in Champaign-Urbana and San Diego, and they are networked with various other mid-level performance centers at other universities.

More recently the NSF made awards for terascale-capacity facilities to the Pittsburgh Supercomputing Center[5] and for the Distributed Terascale Facility[6] (providing tera*grid* capacity) to a consortium including National Center for Supercomputing Applications (NCSA) at the University of Illinois, Urbana-Champaign; the San Diego Supercomputer Center (SDSC) at the University of California, San Diego; Argonne National Laboratory in Argonne, IL; and the Center for Advanced Computing Research (CACR) at the California Institute of Technology in Pasadena. In October 2002, the Pittsburgh Supercomputer Center (PSC) was added to the Terascale Facility.

This evolution of high-performance computing programs at NSF is at the leading edge of evolving architectural diversity in high-capacity computing. In earlier years, the fastest computers used fundamentally faster components (newer technologies, higher cooling and powering,

more complex processor designs). The current state is different – the fastest chips are now also among the most common, and they have very complicated internal structures. Only very specialized problems currently benefit from use of nonstandard parts. (Some of the most technologically impressive processors are found in game machines.) The commercial world continues to demand more computing power, and this huge demand for machines supports investment in new manufacturing processes and designs. High-end computing now depends more strongly on combining very large numbers of these commercially available devices, rather than trying to make unusually fast individual processors.

Parallelism is a recursive notion. Single-chip microprocessors using various forms of internal parallelism are the heart of a computational *node*. For much greater speedup, nodes are combined through switches into physically proximate (to minimize speed-of-light delays) *clusters* of nodes. Now, cluster supercomputers are being distributed over high-speed networks to form *grid computing* environments. The Terascale Initiative is building a large, fast, distributed infrastructure for open scientific research. When completed, the TeraGrid will include 20 teraflops of computing power connected at 40 Gb/s over five geographically distant sites. It will also include facilities for storing and managing nearly 1 petabyte of data, high-resolution visualization environments, and toolkits to support grid computing.

The demand for advanced computing is no longer restricted to a few research groups in a few fields, such as weather prediction and high-energy physics. Advanced computing now pervades scientific and engineering research, including the biological, chemical, social, and environmental sciences. However, the entry barrier continues to be very high. Numerous of our survey respondents observed that, in some areas, the state of the art in computer technology is outpacing tools and best practices from the user perspective. For example, the relatively straightforward and efficient autovectorizing and autoparallelizing compilers of the previous hardware era have given way to complicated messaging directives that must be inserted manually; to many users these are as intimidating and time consuming as programming in assembly language. Industry and academia should work together to remedy this problem and bring greater parity between the available facilities and the tools available for their use.

This issue becomes even more important with the move toward Grid-based capabilities. There is growing mismatch between *theoretical peak* and *actually realized* performance for production codes, as well as a growing investment of time required for users to achieve reasonably good performance. Researchers commented that although the theoretical peak performance of current machines is much higher, they obtain a smaller fraction of theoretical peak today than 10 years ago for many applications. Greater effort is needed to automate the conversion

of code for efficient execution on various machine architectures, including clusters and grids, and to minimize massive code changes as the underlying machines evolve and change.

Many respondents to the Panel's Web survey (details of the survey are in Appendix B) indicated the importance to their work of research-group and departmental-scale computing facilities. We define such facilities as having a factor of 100 to 1000 less capability (e.g., computing, storage) than is provided by the national-scale centers. The proliferation and importance of such resources suggest the need for an effective mechanism – now lacking – to create, nurture, and support them as well as link them into the national cyberinfrastructure. Further, the results suggest that users view national centers as needing to provide capability of order 100 to 1000 times the power of systems generally available to individual academic departments and research groups. Such centers not only dramatically expand the capabilities available to individual projects, but also ensure that all university researchers have equal opportunity. At this point the promise of grids of computers cannot replace the need for both local mid-level facilities and highest-end national resources. Grids are extremely valuable for some types of computations but fail for others because of network latencies and other reasons.

Scientific and engineering applications are covering and will continue to cover even greater time and space scales (e.g., weather, which involves a coupling of scales ranging from planetary waves that last for more than a week, and individual thunderstorms, which are at subcity scale and last for one to a few hours). Such multi-scale problems, often involving the coupling of different models, are exceedingly complex and computationally intensive and thus need sustained high-end computing for the foreseeable future. For example, emerging community climate system models require a sustained 25 teraflops and involve computations closely coupled and thus susceptible to network latencies. But this is only the beginning, as the earth science community moves to comprehensive, high-resolution simulations of combined biological and geoscience models of the environment.

Although many important problems require the highest available processing power, cyberinfrastructure should not concentrate solely on team projects using only the largest and most powerful resources. Rather, it should support a hierarchy spanning a pyramid of machine capacities and the spectrum from small grants to large multidisciplinary centers and projects. As was pointed out in testimony concerning the National Virtual Observatory[10], large team efforts are required to build federations of data and tools to explore them; but smaller groups working independently and given access to these data and tools can (and likely will) make fundamental discoveries.

The current NSF-supported centers remain largely a batch-oriented environment, whereas many future problems will require on-demand

supercomputing for steered calculations and a dynamic environment where the machine needs to respond to the calculation (e.g., dynamic adaptive nesting and the ingest of real-time data that impacts a real-time calculation; such as adaptive sensors in field biology). The current centers are not configured and administered to provide, in most cases, significant fractions of their resources in a dedicated fashion to support the most challenging research problems. Although machines may have the *capacity* to solve huge problems, the users may not have the *capability* to use them effectively because of lack of support for mapping their code efficiently onto specific parallel machines or because of restrictive machine allocation policies.

We received frequent strong input to the effect that the National Resource Allocation Committee (NRAC)[38] allocation process is no longer effective and must be overhauled. For example, users are subjected to double jeopardy by having to prepare both research grant (agency) proposals and proposals for computer resources. Funding of the former with a negative decision for the latter clearly creates a problem. NSF considered coupling the two processes in the early 1990s but chose to leave them separate. Mechanisms for requesting resources should be streamlined as well, and the reviewer base must be broadened to ensure an adequate understanding of the needs being expressed. Moreover, the new allocation process will likely need to include additional types of resources such as federated data repositories and remote instruments.

Even more fundamental issues of resource allocation are intrinsic in cyberinfrastructure concepts of large interoperating grids of computers, instruments, and data repositories. Human committees will not be capable of doing the complex dynamic allocation processes required to balance the supply and demand over thousands of users, hundreds of machines, and numerous variations of computational size and requirements for real-time response. Automated allocation mechanisms are themselves a research challenge and another example of the need for social scientists – in this case economists – to participate.

Both the capacity and demand for high performance computing continue to grow in depth and breadth of use. There continues to be constructive diversity in how this computing is provided and the need for continued experimentation and investment in new machine architecture and supporting software: operating systems, middleware, and application frameworks. On the other hand we need balance (and better yet, real synergy) between extending the frontiers of computing and extending the frontiers of science using computing. The challenge is to both break new ground and bring current and new users along.

A note about sustaining access to highest end computing – The continuing exponential improvement of the hardware underlying cyberinfrastructure provides accelerating opportunities for exercising creativity but can be daunting in terms of managing

the attendant rapid obsolescence of facilities. Maintaining leading-edge cyberinfrastructure requires continuing investment, not one-time purchase. Cyberinfrastructure ("bit-based") investments differ from most other, more "atom-based" kinds. Delaying the start of construction of an accelerator or telescope or research vessel normally increases the cost of the acquisition. Frequently, the opposite is true for computing equipment, which becomes cheaper by waiting a year but becomes obsolete soon thereafter. One way to quantify this is through replacement schedules. Major research equipment may have a realistic lifetime of 10-25 years. The appropriate replacement interval for information technology at the frontiers of performance is closer to 3-5 years. Furthermore, there are changes in the ways machines are used and the types of computations that are needed. As the basic unit costs of information and calculation fall, new ways to get better answers or to displace scientists' time are discovered, and the appropriate levels of local and national computing and the appropriate balance between them will change.

The scientific research world pushes the limits of a number of technologies and acts as a driver for improvement. Collaboration between high end users and commercial providers has been effective and should continue. But the commercial mass markets will continue to determine the computing equipment and services that are most readily available, including the best programming language implementations, fastest chips, and largest disks. The research world has driven very high end networking and the largest computing clusters. There are commercial organizations that specialize in running large computers and disk farms or in taking over entire business functions. They have developed tools and methods for efficient operation to exacting contractual service level agreements, so they provide benchmarks or alternatives for deploying some of the cyberinfrastructure.

3.2 Content

As familiar as the exponential growth in computing, storage, and networking power is the exponential growth in digital information and data. Most all scientific and technical literature is now created in digital form, and large quantities have been converted to digital retrospectively. Scientific, engineering, and medical journal publishing is now done in a hybrid of digital and paper formats with digital taking dominance, although pricing and terms and conditions for use continue as major issues. Some presenters to our panel expressed deep concern about the increasing price of commercially published scientific literature that is forcing academic libraries to collect a smaller and smaller fraction of the overall literature.

The primary access to the latest findings in a growing number of fields is through the Web, then later through classic preprints and

conferences, and only after that through refereed archival papers. The traditional linear, batch processing approach to scholarly communication is changing to a process of continuous refinement as scholars write, review, annotate, and revise in near-real time using the Internet. Major research libraries have switched from microfilm to digitization for both preservation and access.

Crucial data collections in the social, biological, and physical sciences are coming online and becoming remotely accessible; modern genome research would be impossible without such databases, and astronomical research is being similarly redefined through the National Virtual Observatory.[10] Enormous streams of data are arising from observational instruments and computational models. The high energy physics community, for examples, estimates that by about 2012 it will need an exabyte (10^{18} bytes) archive for data from four major large hadron collider (LHC) experiments. The National Center for Atmospheric Research (NCAR)[28] currently has 1000 terabytes of online data and is growing at 10 terabytes per month.

The NSF CISE Directorate through a series of Digital Libraries Initiatives[8, 12] has been instrumental in grounding and informing the emergence of digital libraries in basic computer science and engineering. It has produced important subsystems and institutions and has created synergy among researchers, practitioners (libraries and archivists), and production organizations (libraries, archives, museums). It has enabled research to help define the possibilities, pilot projects to help validate and make concepts real, and partnerships and startups to create new production services. It is a good example of interdisciplinary research, focused by test bed construction, which is needed in a broader cyberinfrastructure program.

A significant need exists in many disciplines for long-term, distributed, and stable data and metadata repositories that institutionalize community data holdings. These repositories should provide tutorials and documents on data format, quality control, interchange formatting, and translation, as well as tools for data preparation, fusion, data mining, knowledge discovery, and visualization. Increasingly powerful data mining techniques are creating greater demand for access to cross-disciplinary data archives. Through data mining new knowledge is being discovered in problem areas never intended at the time of the original data acquisition.

Other trends include the growing need to confederate data from multiple sources and disciplines. The emergence of supercomputing environments capable of executing comprehensive, multilevel simulations (for example, of the environment) requires interoperability between both computational models and the associated observational data from various fields. It was mentioned at a recent meeting of the environmental research and education community that some scientists are spending up to 75% of their time finding and converting data from

other fields. Much of the data being sought is “preserved” in ad hoc and fragmented ways, and all too often ends up in “data mortuaries” rather than archives.

Repeatedly the Panel heard members of the research community citing the need for trusted and enduring organizations to assume the stewardship for scientific data. Stewardship includes ongoing creation and improvement of the metadata (machine-readable and interpretable descriptions of the data itself) by people cross-trained in scientific domains and knowledge management. A key element associated with filling this need is the development of middleware, standard or interoperable formats, and related data storage strategies. Although each discipline is likely best suited to creating and managing such repositories and tools, interoperability with other disciplines is essential, through the creation and adherence to standards, and other means. Additionally, greater emphasis needs to be given to the digitization and stewardship of legacy data (data archeology) and to digital libraries preserving and giving access to past scholarly work.

More and more disciplines are also expressing a compelling need for nearly instantaneous access to databases (both local and distributed) as well as to high-speed streams of near-real-time data from observation and monitoring instruments. Applications such as numerical weather prediction models need to be used in control loops to drive the remote sensors to optimize the data actually being collected; the linkage between data acquisition and processing is now two-way. It is important to note, however, that the technologies for such databases do not yet exist and that many needs of the research community are not accommodated by existing systems (e.g., commercial relational databases). This is a concrete example of how software for large-scale scientific use must extend well beyond the procurement of commercial technology, and often even beyond our current understanding. Thus, both coordinated research into the information technologies and the development of customized technologies for the research community are needed.

Online scientific instruments, or arrays of instruments, are a growing source of digital content for both huge quantities of primary data and the derivative processed datasets. Modern large instruments such as supercolliders and telescopes produce huge streams of data as well as growing numbers of ubiquitous arrays of small sensors. For example, in air and water pollution or seismological monitoring, satellites continue to beam back huge data sets and a growing interdisciplinary community intends to examine practically every aspect of the Earth system from space this decade using data from these satellites.

The emergence of ubiquitous wireless networks offers another big opportunity. Billions of Internet connected cell phones, embedded processors, hand-held devices, sensors, and actuators will lead to radical new applications in biomedicine, transportation, environmental

monitoring, and interpersonal communication and collaboration. The combination of wireless LANs, the third generation of cellular phones, satellites, and the increasing use of unlicensed wireless bands will cover the world with connectivity enabling both scientific research and emergency preparedness to utilize a wide variety of "sensornets". Building on advances in micro-electronic mechanical systems (MEMS) and nanotechnology, smart sensors can be deployed widely, will be capable of multiple types of detection, and can survive for long periods of time[38]. The integration of real-time multisensor data with data mining across large distributed data archives opens further avenues for adaptive monitoring/observation, situational awareness, and emergency response.

3.3 Interaction

We use the term *interaction* in the broad sense of (1) communication between or joint activity involving two or more people and (2) the combined action of two or more entities that affect one another and work together. Higher-performance *computation* provides more powerful tools for discovery through analysis and more systemic and realistic simulations. Acquisition, curation, and ready access to vast and varied types of digital *content* provide the raw ingredients for discovery and dissemination of knowledge. Computation and content, integrated through networking, offer new modes of *interaction* among people, information, computational-based tools/services, and instruments.

Working together in the same time and place continues to be important, but through cyberinfrastructure this can be augmented to enable collaboration between people at different locations, at the same (synchronous) or different (asynchronous) times. The distance dimension can be generalized to include not only geographical but also organizational and/or disciplinary distance. Our surveys confirmed that collaboration among disciplines is increasingly necessary and now requires, in some cases, hundreds of scientists working on a single project around the globe. Cyberinfrastructure should support this type of collaboration in a reliable, flexible, easy-to-use, and cost-effective manner. Groups collaborate across institutions and time zones, sharing data, complementary expertise, ideas, and access to special facilities. This can greatly expand the possibilities for synergy and is especially important to those researchers who are more isolated due to geographic or institutional circumstances.

We also heard that because of converging advances in computation, digital content, and networking, the research community is poised to pursue its work in a much more connected and interactive way. We have the opportunity to extend networked systems to provide *comprehensive* and increasingly *seamless* functional services for research and learning – to create virtual laboratories, research

organizations, indeed technology-enabled research environments that offer a full spectrum of activities in the process of scientific discovery and the education of the next generation. We are at a threshold where a *collaboratory* or *grid community* can become "the place" where a research community interacts with colleagues, data, literature, and observational systems together with very powerful computational models and services. Although many technical, social, and economic challenges remain, the potential exists for facilitating both deeper and broader scientific and engineering research and education.

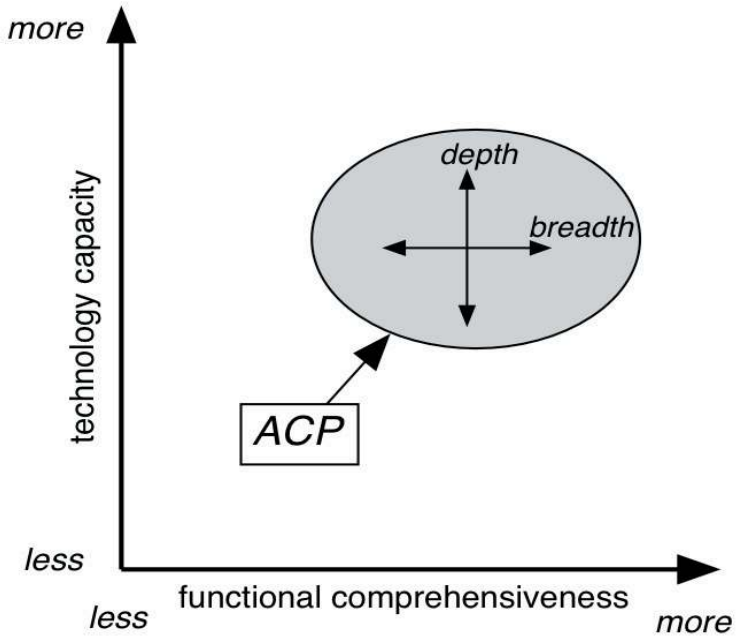

Figure 3.2 – Increasing capacity and functional comprehensiveness of cyberinfrastructure enable both depth and breadth approaches to discovery.

Figure 3.2 is an abstract and qualitative representation of two related dimensions emerging from advances in the nature and application of cyberinfrastructure. The vertical axis is a relative measure of the aggregate basic capability of the technology measured in terms of computation rates, storage capacity, and network bandwidth. The horizontal axis is a measure of breadth of use, or functional comprehensiveness – that is, how completely a cyberinfrastructure-based environment provides the resources and functions that researchers depend upon. To what extent can researchers readily find and effectively interact in a seamless way with all the colleagues, the data, the literature, the appropriate computational services, and the instruments necessary to meet their individual and community aspirations?

Technological capabilities expand rapidly. The Panel also heard, albeit more slowly and less predictably, that cyberinfrastructure is playing a more pervasive role in affecting how scientists do their work. Various fields begin the application of cyberinfrastructure in various ways. For example, some fields are building comprehensive collections of digital science literature; some communities have critical community data repositories and shared libraries of simulation codes; instruments and sensors arrays provide new types of observational data to widely dispersed research teams. The opportunity and challenge are to expand, integrate, and exploit the commonality among these applications of cyberinfrastructure. The shaded area of the graph represents a state of being or state of practice in this cyberinfrastructure capacity vs. comprehensiveness space. The goal of an Advanced Cyberinfrastructure Program (ACP) is to move the state of being region up and to the right (more comprehensive at higher capacity) – both *within* and *among* more and more fields of science and engineering.

As the combined state of capacity and functional comprehensiveness increases, and is adopted more broadly, the payoff will likely derive from enhancing both "depth" and "breadth" approaches to discovery.

In a depth approach, for example, atmospheric scientists could use higher-performance computation (together perhaps with denser and smarter distributed networks of sensors and with higher quality archival data) to improve the resolution and accuracy of a weather prediction model. Astronomers could use a more capable telescope to look more deeply into their favorite region of the universe.

In a breadth approach, a multidisciplinary team of earth scientists could use the availability of more computational power, more complete multi-dimensional data, enhanced observation capability, and more effective remote collaboration services to bring together an entire earth system simulation framework capable of supporting usefully predictive environmental simulations. Astronomers, given access to a federated "digital sky," could explore the breadth of the known universe over the entire available electromagnetic spectrum to seek, for example, rare or new objects or phenomena. We can only begin to glimpse the impact of blended depth and breath approaches, especially as they weave together complementary expertise from multiple disciplines.

Another theme concerning knowledge environments for science based on cyberinfrastructure arose from the testimony we gathered. The theme is one of design of knowledge environments for *multiple uses*. In some cases this means to design such environments with the intent to (at least eventually) support both research and education and build further synergy between them. Others, in a similar context, encouraged intentional activity to use cyberinfrastructure to enhance broader participation ("democratization") in science and engineering. The other variation of a multiple uses environment, sometimes called a *rapid-*

response collaboratory, is to support both basic science and, when necessary, the identification and rapid deployment of scientific and engineering resources to address natural or man made disasters (for example, earthquakes or bioterrorism attacks).

4.0 Achieving the Vision: Organizational Issues

One of the three parts of our charge is to *recommend an implementation plan to enact any changes anticipated in the recommendations for new areas of emphasis*. Our response has been to recommend a major Advanced Cyberinfrastructure Program (ACP) to create, provision, and apply advanced cyberinfrastructure to advance, and ultimately revolutionize, the conduct of scientific and engineering research and allied education. Success for this far-reaching ACP will require synergy among constituencies with varied expertise as well as incentives for participation. The goal of this section is to help NSF leaders create an organizational and leadership structure (some of which, because of its foundation-wide nature, are unusual to NSF) that effectively realizes the goals of the ACP. The Panel has given extensive attention to this part of our charge. We recommend a number of basic principles, processes, and incentives while avoiding being overly prescriptive as to the details so as to allow flexibility for NSF in its implementation of the ACP.

Two complementary activities are to be organized. The first is programs within NSF, which prescribe how resources are allocated to the various activities, evaluate proposals and make awards, and assess outcomes. These programs also represent and advocate for the ACP within the governmental and NSF budget process. The second involves the science and engineering community itself – the researchers, developers, and operational organizations that carry out the missions defined in the ACP. NSF can have significant influence on the organization of the community through setting priorities, defining programs, establishing evaluation criteria for proposals, and then evaluating proposals.

4.1 Elements of the Program

The key elements of the ACP are shown in Figure 4.1. The proximate outcome is new ways of conducting research through the application of information technology. The conduct of science and engineering research is built (in part) on these applications, which are tailored to the specific needs of people, groups, organizations, and communities conducting that research. Thus, the ACP directly funds activities resulting in the conceptualization, implementation, and use of such applications—it is not focused on cyberinfrastructure alone. Some applications are generic (such as distributed collaboration), and many others are discipline specific (like distributed community access to a specific scientific instrument).

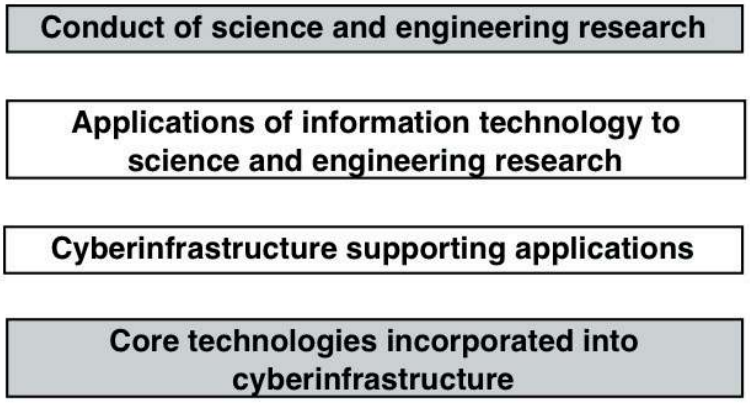

Figure 4.1. A layered architectural view of the ACP. The shaded boxes fall outside the scope of this report.

Applications are enabled and supported by the cyberinfrastructure, which incorporates a set of equipment, facilities, tools, software, and services. The ACP supports the creation and operation of advanced infrastructure tailored to specific domains, but it obviously does not include the core funding for the research (the top shaded box). Likewise, the ACP includes support for research on systems issues relevant to bringing together a heterogeneous mix of technologies (hardware, software, communications, storage, processing) to support advanced applications. Core technologies in the lower shaded box encompass the bulk of the current CISE research budget and should be preserved rather than reallocated to the ACP.

4.2 Technology Research and Technology Transfer

While the ACP is about revolutionizing the conduct of research, an equally important opportunity is to transform information technology itself. To illustrate this important aspect of the ACP, a second technology-transfer dimension is added in Figure 4.2. The three major phases of technology transfer (further elaborated and subdivided in Appendix C) are *applied research* (conceptualizing and bringing new application and infrastructure ideas to fruition), *development* (creating new technology artifacts ready for deployment), and *operations* (installing these software artifacts and enabling facilities and equipment, integration, keeping them running, and supporting end users). These phases are all relevant to both applications and cyberinfrastructure.

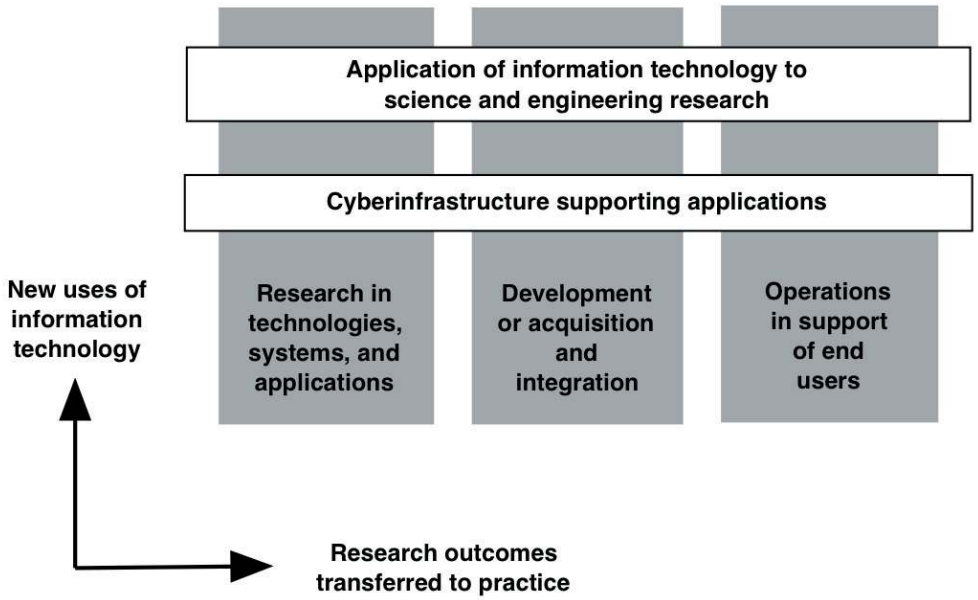

Figure 4.2. Technology transfer adds another dimension, where operations are supported by development, which is based on research outcomes.

4.3 Some Challenges

This ACP is ambitious, and as a starting point for considering its organization, we must recognize the most serious challenges to its success.

Only domain scientists and engineers can revolutionize their own fields. At its core the ACP involves rethinking the processes and methodologies underlying individual scientific and engineering fields. Domain scientific and engineering researchers must step up and enthusiastically create and pursue a vision.

Computer scientists (and allied technological fields, such as information science, and electrical engineering) must be involved. The substantial and ongoing involvement of information technology specialists is required to ensure that innovative new uses of technologies are identified, existing technologies are molded in new ways, and research into new technologies and new applications of technology is informed by opportunities and experiences in science and engineering research.

Taken together, these two issues present a serious challenge to any organizational structure. If the organization is weighted too heavily toward the domain scientists, the focus overemphasizes procurement of existing technologies, and computer scientists become viewed

as “merely” consultants and implementers. If the weight shifts too heavily toward computer science, the needs of end users may not be sufficiently addressed, or effort shifts too heavily toward creating new technologies with insufficient attention to stability and user support.

Commonalities across science and engineering disciplines must be captured. Absent appropriate levels of coordination and sharing of facilities and expertise, there would be considerable duplication of effort, inefficiency, and excess costs.

Collaboration across science and engineering disciplines must be empowered and enabled, not impeded. Too often information technology becomes a source of Balkanization and an obstacle to collaboration or innovative change. The goal of the ACP is to make the cyberinfrastructure and applications an enabler (not an obstacle) to opportunistic and unanticipated forms of collaboration across disciplines, as well as encourage the natural formation of new disciplines. As in achieving commonalities, realizing this goal requires a largely collective effort.

Social scientists must work constructively with scientists and technologists. The social scientists can assist in understanding social and cultural issues underlying the direction of the ACP and, like technologists, can aid research in their own disciplines based on the experience gained.

4.4 Organization within NSF

The ACP will be retrofitted to an NSF organization whose primary mission, the conduct of science and engineering research and education, remains unchanged. It will be important and challenging to pursue major changes in the organization and processes underlying NSF’s primary missions to promote innovative application of information technologies, while avoiding significant organizational disruptions. Thus, we suggest that the organization of the ACP be overlaid in a matrix fashion on the existing organizational structures with the addition of a new coordinating ACP Office (ACPO).

As a starting point, the structure of Figure 4.1 is modified to align better with the research disciplines represented at NSF and becomes Figure 4.3.

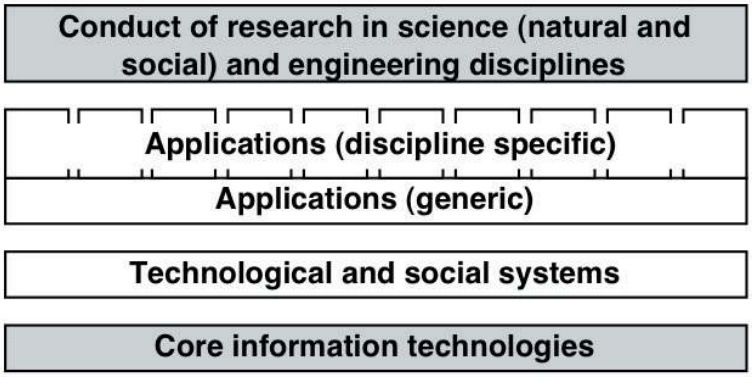

Figure 4.3. Relationship of the layers of Figure 4.1 to underlying disciplines. Applications are a hybrid case with shared responsibility between technological and disciplinary programs.

Cyberinfrastructure brings together many technologies (hardware, software, processing, storage, communication, etc.) to provide a coherent end-to-end functionality in support of applications; that is, at its heart cyberinfrastructure is a *technological system*. Many core technologies have themselves a system flavor, but we distinguish technological systems at the top level of hierarchy—where technology meets applications and uses—and observe that systems in this sense have special significance to both cyberinfrastructure and to applications. Figure 4.3 also emphasizes that, in the context of the fundamentally social enterprise of science and engineering research, technological systems as defined here and social systems (groups, organizations, and communities) are fundamentally intertwined.

Insofar as possible, applications should be generic, seeking to serve a variety of disciplines, but with sufficient flexibility, configurability, and extendibility to accommodate local variations and extensions. This contributes to both commonality (enabling future cross-discipline collaboration) and efficiency (through sharing of resources and expertise). On the other hand, there are clearly discipline-specific needs as well, with many organizational and process changes not readily transferred to other disciplines. A common cyberinfrastructure encourages commonalities and opens the door to future cross-disciplinary collaboration.

The organization within NSF should mirror the types of players (deliverers of research, development, and operations) illustrated in Figure 4.4.

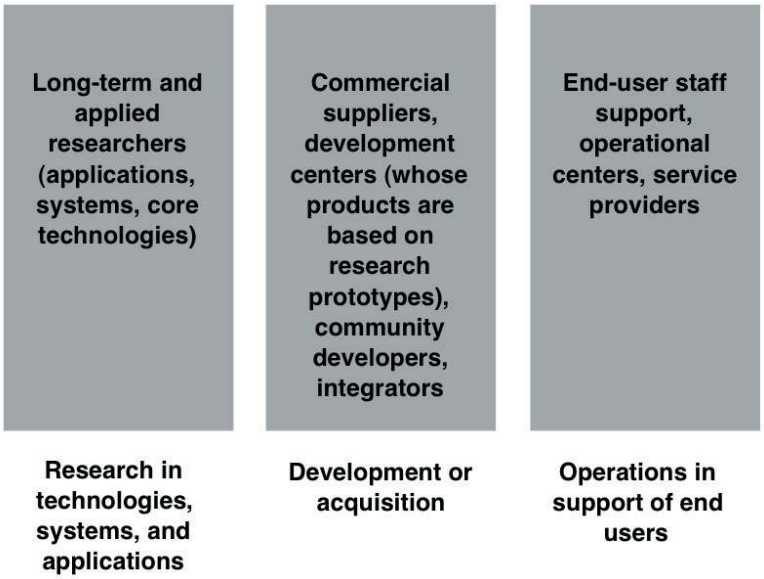

Figure 4.4. Summary of specific players delivering parts of the ACP

In terms of internal organization, our proposed division of responsibility is illustrated in Figure 4.5 (for applications) and Figure 4.6 (for cyberinfrastructure). We envision the initiative being served by a matrix management structure with the direct involvement of all of the NSF directorates. Some overriding principles (referencing Figures 4.5 and 4.6) can be stated.

Domain science and engineering directorates must take the lead in revolutionizing their respective fields through new research organization and processes, supported by new applications of information technology. We envision a program in each interested directorate (and we hope they will all be interested) that takes primary responsibility for formulating and implementing a vision, fostering buy-in and participation of its respective scientific or engineering research community, and creating a coherent program. Such efforts need to be open and oriented toward mutual coordination among directorates and should emphasize common standards and employ a common cyberinfrastructure.

CISE must be deeply involved both in serving as a technology leader for the overall initiative and in using scientific applications and experience of application users to inform its own technology research. CISE should be primarily responsible for both cyberinfrastructure and generic applications (much as it has managed the PACI program in the past) while also improving specific areas as outlined in Section 5. A primary goal of cyberinfrastructure is to capture the major technology requirements and provide tools to aid

in application development, thus minimizing the need for technology-specific activities in other directorates. CISE will take responsibility for identifying commonalities among the needs of different disciplines. It should also lead the effort to define common infrastructure and standards that ensure that commonalties are captured and that future interdisciplinary collaboration is encouraged. CISE should be responsible for ensuring that the ACP is founded on a vibrant research agenda in technological systems and applications and that the research feeds the development of prototypes, production services, and commercially valuable end products. Finally, CISE should include and cooperate with SBE in conducting underlying research in the social aspects of both systems and applications.

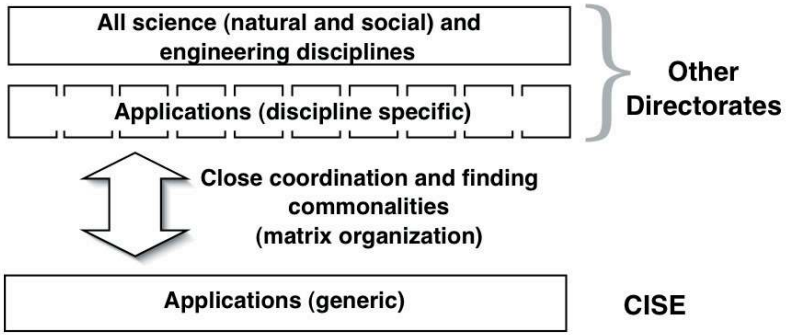

Figure 4.5. Assignment of responsibility for the vision and governance of applications to the NSF directorates.

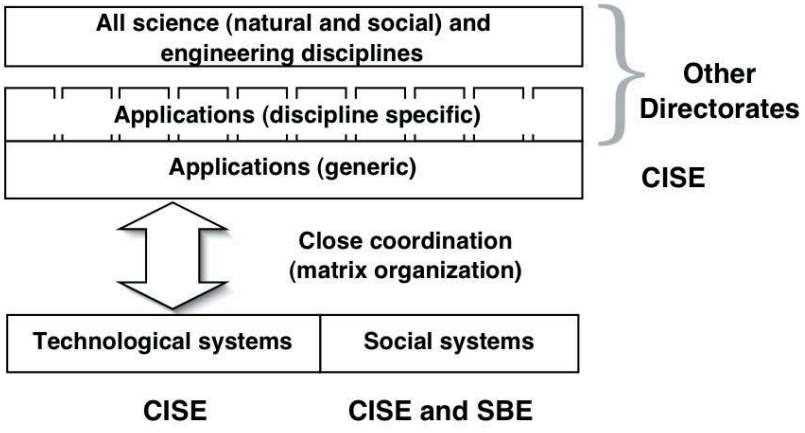

Figure 4.6. Assignment of responsibility for the vision and governance of cyberinfrastructure to NSF directorates.

To meet the challenges of achieving commonalities and collaboration, it is critical that the constituent programs within each directorate be viewed as parts of a foundation-wide initiative, while seeking to ensure that each respective community is served well.

Maintenance of sufficient coordination within the proposed matrix management structure will be formidable. **We therefore recommend that a single coordinating ACP Office (ACPO) be established to provide overall vision and guidance and exercise budgetary planning and responsibility.** (This office may or may not be an "Office" in the usual NSF meaning of the word. It could be administratively hosted in CISE or elsewhere, but it needs significant autonomy as described in this section.) The ACPO defines the overall vision of the ACP and represents and advocates this vision internally and externally to NSF. It develops budgets for the ACP, including the overall budget and sub-budgets for the various activities and the directorates. It serves as a central point of coordination among the complementary activities, including the identification and pursuit of commonalities and achieving uniformity and consistency where appropriate.

The directorates are the primary source of vision for their respective disciplines, and they formulate proposals to the ACPO for new programs and solicitations, insofar as appropriate in collaboration with (and as appropriate jointly with) other directorates. The ACPO evaluates the merit of those proposals consonant with the coordinated direction of the ACP, including assessing past efforts and seeking advice from the community. The ACPO then determines (or at least recommends to the Director of the Foundation) budgetary allocations to the various directorates based solely on the merit of their proposals and an evaluation of how these pieces fit together constructively in the overall coordinated activity of the ACP. The ACPO also represents the ACP in coordination with various other agencies and international bodies.

It is important that the ACPO view itself as the leader of a revolution in the conduct of research, and not primarily as an "information systems" or "information technology procurement" organization (a common organizational construct in government and industry).

The ACPO will not directly evaluate or fund projects in the community, this being the responsibility of the individual directorates. The reporting relationship of the ACPO should maintain budgetary independence from other programs in the directorates and place the ACPO in a position to strongly represent the budgetary needs of the ACPO within NSF and the government.

The leader of the ACPO is an especially important responsibility. **Its leader must have fundamental responsibility for achieving these goals, with sufficient credibility, power, and authority to succeed**.

This highly qualified person should be visible and highly placed, able to lead a large and complex matrixed operation with a substantial budget. Whether this leader is a discipline or computer scientist or engineer is secondary (a broad background and interests is ideal); most important is that he or she be deeply committed to successfully achieving the vision of a revolution in science and engineering research and be willing to explore and learn in the process.

The leader of the ACPO, although perhaps attached to an existing directorate, should be a functional peer with the assistant directors of the NSF directorates. A position at this high level is necessary to attract the right combination of visionary and manager, and to represent NSF as the leading U.S. agency in cyberinfrastructure when dealing with other federal agencies and international partners. An example of a structure that can be considered is as follows:

- The leader of the ACPO would report to an ACP Steering Committee consisting of the assistant directors of all involved directorates and chaired by the CISE AD (in recognition of the special role of CISE in the ACP).

- The Steering Committee would meet regularly with the leader of the ACPO and assume collective responsibility for the success of the initiative. The leader, working with the Steering Committee, would be delegated primary responsibility over a budget allocated to the ACPO.

- The ACPO leader would work with the Steering Committee in program generation, allocation of budget to directorates, awards, and oversight. The appropriate directorates working with their respective communities would carry out the details of this work.

- The leader of the ACPO would also be responsible for NSF liaison to other relevant programs in federal agencies and international bodies.

- The ACPO is intended to be the coordinator of an effective matrix organization, not a large organization duplicating or replacing the normal directorate activities. The ACPO would have a modest staff to help in budget and program development and performance reviews.

Appendix C includes more discussion of roles and organizational options.

4.5 Organization of the Community

Much of the work of the ACP will be carried out by individual research groups in the science and engineering research community, who will provide vision and experimentation and who will ultimately conduct

research in new ways. The development and integration portion of the ACP, as well as operations of a common infrastructure and centralized user support, will be carried out by new or existing centers in the community funded by NSF. These centers are divided into several categories, including development centers, generic centers serving the science and engineering communities broadly, and disciplinary centers.

The ACP requires an organization for internal NSF coordination, as well as a central point of coordination in its external implementation. One or more development centers should be devoted to activities at the core of the initiative. These core activities include the planning, acquisition, integration, and support of the major software platforms and components at the foundation of the cyberinfrastructure. This includes choice of commercial software for underlying computing platforms and, where available and appropriate, for middleware and application components. These core activities will call for close coordination with industry, including the possible use of industrial products or prototypes as a basis of the ACP, and assistance with the transfer of successful technologies developed within the initiative into commercialization. Other core activities include the productizing of research prototypes, the development of new capabilities, and the integration of all these elements into a uniform software release with subsequent maintenance, support, and upgrade. In some cases, software may be maintained and upgraded by the community (e.g., open source), in which case the core activity includes governance of the process, such as choosing patches or upgrades to include in the releases. Development centers may be contracted industrial firms, existing laboratories, or new centers set up for this expressed purpose. The ongoing NSF Middleware Initiative provides (on a smaller scale) valuable experience and guidance in the organization of this portion of the ACP.

Generic centers focus on operations and user support for applications and infrastructure serving the broad research community, and discipline centers focus on applications and infrastructure more specialized and dedicated to particular disciplines, and include strong expertise in a discipline and its particular needs and challenges. Generic centers are needed to pursue broad commonalities, while disciplinary centers can accumulate disciplinary skills and thereby better meet specific disciplines needs.

There is no intention that these activities be strongly separated; development, generic, and disciplinary activities may be co-located or even grouped within common centers. One appealing organizational model, for example, is a development or generic center that maintains and integrates a collection of disciplinary groups.

Processes - As emphasized in Figure 4.2, several distinct activities each make essential contributions to the ACP. One such contribution

is research—a traditional emphasis of the NSF—but there are others, broadly defined as development, operations, and use. These are decidedly not independent activities. Technology transfer seeks to benefit science and engineering research by employing the best ideas arising from research. But research agendas also should be influenced by the vision for the future conduct of science and engineering research. Similarly, there is a vertical flow of ideas and influence. Applications are influenced by emerging or anticipated capabilities in cyberinfrastructure, which are influenced in turn by advances in core technologies. And core technology research should be informed by anticipated cyberinfrastructure requirements, which in turn is influenced by capturing commonalities among application opportunities.

The research supporting applications in Figure 4.2 will increase the collaboration among computer scientists (and related disciplines, such as information science and electrical engineering) and domain scientific and engineering researchers (including the social sciences) to the benefit of all sides. Similarly the research supporting technological and social systems will increase the visibility of research into information technology systems in the broad sense, incorporating processing, storage, and communication into holistic social-technical systems solutions.

It is informative to examine the internal organization of the CISE directorate in light of these changing and magnified responsibilities. The vertical organizational structure of Figure 4.3 would focus attention most squarely on the greatest challenges mentioned earlier and highlight research into systems and applications. However, care should also be exercised that research efforts devoted to advancing core technologies receive continued high priority, as these efforts remain a critical underpinning of both the ACP and the nation's industry and economy.

Following the successful Internet experience and the more recent NSF middleware initiative, we expect that the development process leading to structure shown in Figure 4.2 will focus on the productizing and integration of a combination of commercially available software and research prototypes. The ACP must maintain a balance between deploying and gaining experience with emerging technologies, while providing users with a stable environment that is well documented and supported. The goal of development is thus to create and evolve a unified software distribution that is well maintained and supported. Of course, the development and operations are undertaken by experienced organizations funded by NSF, normally under cooperative agreements. The longer-term goal should be the commercialization of successful cyberinfrastructure and applications, with NSF continuing to fund development at the frontiers of noncommercially available solutions.

The operations stage will mix two models, as appropriate: a software distribution that can be installed, operated, and supported within the end-user organizational context, and software that is centrally operated to provide services over the network. NSF will fund organizations prepared to develop, maintain, and upgrade software distributions made available to end-user organizations and also organizations that operate cyberinfrastructure and/or applications provided as services invoked over the network. A proper and evolving balance should be maintained between professional staff supporting centralized operations and end-user operations, taking into account tradeoffs between the greater accountability and familiarity of local staff versus the efficiency and sharing of resources and expertise arising from centralization.

Incentives - The three primary activities identified in Figure 4.2 have very different metrics for evaluating proposals and outcomes.

Research is a competition of ideas. Allocation of resources starts with the program announcement and evaluation of the resulting proposals. This is bottom-up, stating the evaluation criteria with detailed initiatives arising from the research community. Overlap or duplication is acceptable where different researchers pursue competing visions for accomplishing similar ends. Post-evaluation is based on the intellectual quality and impact of the research outcomes.

Development is a competition of plans. An overriding goal of development is to limit duplication of effort, and concentrate resources on a set of integrated and maintained software distributions collectively covering the scope of the ACP. Thus, development is partitioned and assigned to organizations based on the responsiveness to needs and credibility of their plan for pre-defined concrete outcomes. Post-evaluation is based on how effectively the plan has been implemented and also on how extensively the outcomes are adopted and used and on user satisfaction.

Operations is a competition for users. Operations serve end-users, domain scientists, and engineering researchers, responsively providing service and support. There should be two or more competitive operational options available to users. A primary point of post-evaluation should be the satisfaction of the users who are served, and to a lesser extent the number of users who are served, based on input from the user community.

These distinct evaluation criteria should not suggest that these activities must be strongly separated organizationally; to the contrary, there may be advantages to grouping applied research, development, and operations (or some subset of these activities) within a common organization and geographic location.

Continuity - Human resources are critical to getting cyberinfrastructure and applications working, keeping them working, and providing user support. In the interest of funding more grants, NSF has arguably undersupported the recurring costs of permanent staff, preferring to focus resources on acquiring “hard” or “tangible” assets or the support direct research costs. In the ACP, human resources are the *primary* requirement in both development and operations, and success is clearly dependent on adequate funding both in centers and in end-user research groups.

Where possible, off-the-shelf commercial technologies and services should be acquired, but advanced and experimental capabilities will require NSF support of applied research, development, and operations. Success depends on specialized skills not readily available in the job market; rather, the most valuable staff will arrive with generalized programming and system administration skills and then learn valuable specialized skills through years on the job. A starting assumption in the funding of development and operations organizations should be continuity and long-term commitment. Absent significant problems and negative evaluations, funding initiatives in these areas should work from a base assumption of at least a ten-year lifetime for each participating organization. This is not to minimize the importance of ongoing evaluation and feedback, nor is it intended to preclude the redirection of funding from poorly performing organizations.

5.0 Partnerships for Advanced Computational Infrastructure: Past and Future Roles

5.1 The Past and Present

We have described prior NSF sponsored investments that have collectively created a platform for major science-driven expeditions to develop and apply advanced cyperinfrastructure. The longest running and largest investment has been a series of initiatives to advance U.S. science and engineering by providing computational resources, including the Partnerships for Advanced Computational Infrastructure (PACI) program. Our charge specifically asks us to assess the effectiveness of the PACI program and to make recommendations about its future in the context of any new directions we propose.

Advanced computing programs began in the early 1980s, when the most powerful machines at that time —"supercomputers"— were not available to the entire U.S. scientific community. Hence the predominant need was for access to computing cycles at the highest end, and as a result five NSF Supercomputer Centers were founded in 1986 and 1987. The PACI program, established in 1997, was the next step. The goals of the two PACI partnerships (hereafter called "the PACIs") — the National Partnership for Advanced Computational Infrastructure (NPACI) and the National Computational Science Alliance (hereafter called "the Alliance") — were much broader than furnishing access to high-end compute power and the associated services. Their missions included provision of data storage and networking, education and outreach, and fostering of interdisciplinary research. At the center of each PACI partnership is a leading-edge site — the National Center for Supercomputing Applications (NCSA) for the Alliance, and the San Diego Supercomputer Center (SDSC) for NPACI. The PACI program is explicitly not allowed to support basic research.

Following the guidelines of the original PACI solicitation, the activities of the PACI partnerships have addressed multiple needs and served multiple purposes, some of which we highlight:

- During the five years of the current program, the two PACI partnerships have fulfilled their mission of providing high-end computing cycles. This conclusion is based on systematic, regularly conducted user surveys that are reported to NSF, and on the survey conducted as part of this panel's information-gathering process (Appendix B).

- The PACIs have supported, engendered, and supplied software tools to help users take advantage of architecturally diverse, increasingly complex, and distributed hardware. In addition to

joining and enhancing pre-existing software activities such as Globus[40] and Condor[41], the PACIs have initiated diverse projects involving all aspects of high-end computing. Two examples are the Access Grid[36], used at more than 100 sites worldwide, and the Cactus[42] programming framework, an open-source environment that enables parallel computation on different architectures along with collaborative code development.

- Through a joint Education, Outreach, and Training[32] activity, the PACIs have broadened access to computational science and engineering by encouraging the participation of women and under-represented groups at all educational levels.

- Many successes in domain science and engineering have been enabled as well as supported in part by PACI funding. In particular, some PACI-enabled collaborations among domain scientists and computer scientists have been exemplars of interdisciplinary interactions in which information technology becomes a creative, close partner with science. To name one among many, the recently funded National Virtual Observatory[10] which includes participants from the Alliance and NPACI, was described as a top priority in the 2001 U.S. National Academy of Sciences decadal survey of astronomy and astrophysics[43]. To a degree beyond anything anticipated even five years earlier, the National Virtual Observatory links astronomy with cyberinfrastructure in the forms of grid computing and federated access to massive data collections. The National Virtual Observatory concept grew from collaborations associated with the PACI program, and illustrates how advances in computer science and information technology can inspire new methodologies and directions in science, not just traditional science that is bigger and faster.

- International collaboration is an inherent part of computational science and engineering, and the PACIs are regularly involved with leading international consortia such as the Global Grid Forum.[22] Individual scientists supported in part by PACI are leaders in visible international projects such as GridLab[44], which involves Grid computing and numerical relativity.

The PACI partnerships have been reviewed annually by a program review panel convened by NSF. These reviews have been consistently positive with respect to the overall achievements of the Alliance and NPACI as defined by the criteria of the PACI program. However, not surprisingly for such a large and complex program, different aspects of the program have had different degrees of success. This is not meant as a criticism; it would be unrealistic to expect perfection in every element of the PACI program, which created new organizations with notable differences from the original supercomputer centers mentioned earlier.

Turning now to issues of concern, the PACI program has exhibited, from its beginning, a tension between two needs that cannot easily be reconciled: providing production systems for the current generation of high-end users, and moving to the next highest level of computing capability. Since the program's core funding has never been adequate to support more than one generation of computer system, tradeoffs have been inevitable.

In addition, the annual program review panels have expressed repeated concerns about the overall effectiveness and responsiveness of PACI activities in discipline-specific codes and infrastructure ("application technologies") and, to a lesser extent, generic software and infrastructure for high-end computing ("enabling technologies"). We discuss these concerns further below.

The PACIs have unquestionably had significant success and impact. Nonetheless, we believe that certain changes, described in the next section, should take place so that the PACIs, or their successors, become an integral part of the ACP proposed here.

5.2 Rationale for the Future

Part of the charge to the present panel was to evaluate the performance of the PACI program in meeting the needs of the scientific and engineering research communities. Given our broad definition of cyberinfrastructure we have interpreted this charge as an opportunity to consider potential roles for the PACI partnerships in a greatly expanded context. Since the Pittsburgh Supercomputing Center (PSC)[5] was selected by NSF in 2000 as the site for the Terascale Computing System[45], we include PSC as well as the PACIs in our discussion of the future.

The panel believes that today's science and engineering research continues to require computing resources at ever-higher levels and in ever-wider dimensions.

- The need remains, exactly as described in the 1995 Hayes Report[36], for the U.S. science and engineering research community to have access to machines that are substantially more powerful than those available at typical research universities, and for support services to enable those machines to be used most effectively.

- We anticipate increasing demand for advanced networking capabilities (including speed, bandwidth, quality of service, and security) for the indefinite future.

- The importance of data in science and engineering continues on a path of exponential growth; some even assert that the leading science driver of high-end computing will soon be data rather than

processing cycles. Thus it is crucial to provide major new resources for handling and understanding data; the National Virtual Observatory (briefly described in Section 5.1) emerged from recognition that the data avalanche in astronomy requires digital archives, metadata management tools, data discovery tools, and adaptable programming interfaces.

- Finally, sustained work is needed on software tools and infrastructure that enable general use of computing at the highest end, as well as on discipline-specific codes and infrastructure. It is universally agreed that producing and maintaining widely usable, reliable software is at least one, possibly several, orders of magnitude more difficult than generating an initial high-quality prototype.

As described in Section 2, the Panel is recommending a broad Advanced Cyberinfrastructure Program whose goal is to transform the conduct of science and engineering research, and which includes significant, sustained new funding for both discipline-specific and generic enabling infrastructure. Since the ultimate drivers of cyberinfrastructure are the needs of the scientific and engineering research communities, we believe strongly that those needs will be addressed most effectively by ensuring that enabling and application infrastructure projects associated with the ACP receive rigorous peer review. This is a fundamental change from the all-in-one structure of a PACI partnership, whose activities have been funded and reviewed as a unit. Our view is based on both philosophical and practical reasons.

Organizationally, this would be accomplished by creating new applications-focused programs within each interested NSF Directorate, as discussed in Section 4. These programs would also create any discipline-specific cyberinfrastructure required to support these applications, often based on extensions to the more generic cyberinfrastructure. Each of these programs would seek to create and execute a broad vision for revolutionizing research within their respective disciplines through the support of peer-reviewed projects. In many cases, we expect participation in these projects by the PACIs and other ACP-supported centers in partnership with disciplinary experts. The justification for this is the belief that disciplinary experts, in close partnership with computer scientists, are best able to judge the merits, impact, and importance of applications and specialized cyberinfrastructure focused on their field, and that these projects should be peer reviewed rather than initiated by the centers. In addition, reviewers who have substantial experience with software development, who take a broad view of high-end computing, and who will pay attention to opportunities for complementary activities and unnecessary duplication, should assess the quality of cyberinfrastructure projects.

The practical motivation for recommending separate peer review of application and enabling technology activities rests on the following observations, frequently made during the panel's information-gathering phase:

- The PACIs are not standalone, but partnerships involving many partners. Commitments have been made, explicitly or implicitly, to a number of partners, and these partners are represented in the PACI management structure. Thus it is difficult to phase out activities of existing partners or add new partners.

- There has been only limited review of enabling and application technology activities, particularly in assessing their impact on the relevant users and communities.

The peer review process that we envisage must always include consideration of the quality of each proposal's computer science and information technology aspects. To be specific, infrastructure projects in application areas need to be peer-reviewed by both domain and computer scientists, as are the current Information Technology Research (ITR)[46] proposals, to assess their quality based on criteria defined by the needs of cyberinfrastructure for the particular scientific community. In this regard, it is important that there should be no artificial distinction, as there was in the original PACI program, between research and development; the best enabling and application infrastructure projects, almost without exception, include both. Enabling and application infrastructure projects can be proposed by researchers and teams from any eligible institution or group of institutions, including, of course, the current PACI leading-edge sites and/or their partners. It is essential, however, that non-PACI teams to be given an opportunity to compete for this funding.

It is entirely consistent to believe, as the panel does, that the PACI program has had many successes, and at the same time to recommend a new structure for the future. We repeat our awareness of the outstanding results that have been achieved in both application and enabling technologies by PACI-supported efforts. In no sense are we advocating that such efforts be curtailed; in fact, our expectation is the opposite. Given the expertise developed at leading-edge PACI sites, proposals involving these groups should have a high success rate in peer-reviewed settings. Peer review of application and enabling infrastructure projects is therefore unlikely to be harmful to the best teams currently supported by PACI funding, while opening funding opportunities to a wider field.

5.3 The Future of the PACI program

To preserve the many accomplishments and talented personnel associated with the PACI program while the ACP is being defined, the panel recommends a two-year extension of the current PACI cooperative agreements. After those two years, until the end of the original ten-year lifetime of the PACI program, the panel believes that the two existing leading-edge sites (NCSA and SDSC) and PSC

should continue to be assured of stable, protected funding to provide the highest-end computing resources. In addition, the two PACI partnerships should continue their activities in education, training, and outreach. At the end of this period, there should be another competition for the roles of "leading-edge sites", possibly renamed, with (if appropriate) revised missions and structures.

Based on the assumption that sufficient new funding is in place, the new, separately peer-reviewed enabling and application infrastructure part of the ACP would begin in 2004 or 2005, after the two-year extension of the current cooperative agreements. New funding is absolutely essential to retain experienced PACI staff and to maintain already-established successful collaborations in enabling and application technologies. As observed in Section 4, *trained and knowledgeable people are the single most important component of cyberinfrastructure.*

With this timeline – a two-year extension of the current agreements and a major infusion of new funding in 2004 of 2005 for separately funded, peer-reviewed infrastructure projects – coupled with a partial disaggregation of functions through 2007, the panel believes that stability will be ensured for parts of the PACI program where it is most needed. Our further hope is that this schedule will reduce the energy and anxiety associated with submission of the annual program plan.

6.0 Budget Recommendations

6.1 Scope of the Program

Achieving the vision of the Advanced Cyberinfrastructure Program (ACP) will require coordinated NSF support of a broader set of activities and facilities than the agency has historically supported. In addition, existing activities (e.g. providing access to high-end computers, enduring data archives, and middleware software development) will need substantially higher funding levels. NSF's role is not limited to financial backing — it is also critical that NSF provide an effective organizational structure to coordinate the ACP, establish operational and user support centers, provide leadership for the nation (including other research funding agencies), and coordinate with similar international activities. This requires not a one-time or short-term initiative, but rather the Panel advocates a material modification to the direction and priorities for the Foundation through a program of sustained long-term funding. In this section, we provide our best estimates of the level and allocation of funding needed near the beginning of the ACP, although we expect this estimate to be modified over time as needs and priorities change.

As described in Section 2, information technology tools and resources should not only support high-end numerical simulations and network connectivity (the major emphases in the past), but also digital libraries, instruments for data acquisition, massive archives of observational data, community application frameworks, and collaboration tools for routine use by researchers. Research communities and disciplines should be able to prototype, refine, develop, and deploy community-specific distributed applications. Robust software (both cyberinfrastructure and application) must be developed, maintained, upgraded, distributed, supported, and in some cases (as in distributed middleware, data curation, and scientific computing) professionally operated. To make these tools and resources accessible across a wide range of academic institutions, we must create a "grid" that provides convenient access to distributed resources, both in the United States and internationally.

Two very important principles that the Panel would like to maintain are:

- The high-end scientific computational resources available to the United States academic research community should be second to none.

- NSF, in collaboration with other appropriate mission agencies, should take lead responsibility for creating and maintaining the crucial data repositories necessary for contemporary, data driven science. The definition of "crucial" will come from the research communities.

The resulting cyberinfrastructure will be much more comprehensive in function and scope, and will be utilized by many more researchers than past NSF infrastructure programs (with the possible exception of the Internet). To gain maximum benefit, it is crucial that NSF support not only the development, provisioning, and operation of cyberinfrastructure and applications, but also their use in the daily conduct of science and engineering research. While support of domain science and engineering research per se is outside the scope of the ACP, successful use in the conduct of this research does require adequate professional staff to provide advice, assistance, and technical support, and these services are within the scope of the ACP.

To achieve the greatest benefits and broadest use, and also to work against Balkanization that inhibits interdisciplinary collaboration, commonality must be captured across disciplines, solutions for common issues identified and solved, and interoperability facilitated through standardization and the choice of common technical solutions. This is another important budgetary priority for the ACP.

The charge to this Panel included the request to "recommend an implementation plan to enact any changes anticipated in the recommendations for new areas of emphasis." In the following, the budget requirements of this broad spectrum of activities are estimated. In the course of describing these needs, we supply additional recommendations and detail an implementation plan.

6.2 Budget Summary

A high-level summary of the budget is given in the following table. Later subsections describe each of these activities in greater detail.

Estimated annual budget	Millions of $ per year	
	Subcategories	**Total**
Fundamental and applied research to advance cyberinfrastructure		**$60**
Research into applications of information technology to advance science and engineering research		**$100**
Acquisition and development of cyberinfrastructure and applications		**$200**
Provisioning and operations of cyberinfrastructure and applications		**$660**
Computational centers	$375	
Data repositories	$185	
Digital libraries	$30	
Networking and connections	$60	
Application service centers	$10	
Total		**$1020**

These amounts are meant to be in addition to the current NSF investments in these areas, with the exception of the $375 million per year for "computational centers," which does include the current level of funding of approximately $75M/year. The funding described here augments, leverages, and creates incentives for exploiting commonality in the cyberinfrastructure investments already underway in the various NSF directorates. These funding recommendations are for NSF programs only, and presume that other federal agencies and institutions will continue to invest in related research and development. The ACP would increase its funding level as the program is defined and implemented. We estimate a credible ramp up to $545M/year of additional funding over two years and to the full $1020M funding in three years.

6.3 Discussion of Budget Categories

This section provides additional information and justification for the budget estimates. Our primary methodology was to estimate, in each category, how many individual projects and centers meet the goals of the program, and what average level of funding would be appropriate for each in order to reach a desirable critical mass. These "per project/center" costs are estimates and are average annual budgets, not upper bounds, and we would expect a range of actual expenditures around this average.

When budgetary components such as centers, research activities, equipment, and data repositories are described separately, they are not necessarily meant to be freestanding entities. They are elements of one overarching integrated, multidisciplinary, systemic program. It would be appropriate to co-locate and put some of these under a common management umbrella, thus benefiting from increased economies of scale and aiding overall coordination. For example, disciplinary-based data coordination projects may be affiliated with one of the data repositories, and large-scale operational centers may house substantial software development and deployment projects. In addition, we sometimes describe projects, and these may or may not be organizationally located within centers.

While we use the number of centers as one element of a budgetary estimate, the Panel generally provides a range rather than advocating a single hard number. Details at this level should be based on substantial analysis and community input, taking into account a number of factors, including existing resources, economies of scale and scope, availability of appropriate sites and institutions, and the willingness and ability of the community to establish and manage such activities. The actual outcome may reasonably differ materially from our recommendations.

The Panel does, however, feel strongly about several points:

- The existing centers (the leading-edge sites for the Alliance and NPACI plus PSC, and perhaps NCAR) have already accumulated significant expertise and experience relative to the ACP, and, subject to appropriate reviews, are likely to be among the initial sites;

- The supporting systems (data storage, high-performance computers, networks etc.) made available to United States academic researchers should be second to none, and

- There should be sufficient capability (scientific application performance, memory size, I/O speed, etc.) and available job time on such systems to support dozens of qualified groups conducting high-quality and high-impact research utilizing these systems.

Each of the major budget categories will now be discussed further.

Research to advance cyberinfrastructure - As discussed in Section 4, cyberinfrastructure is a system incorporating many processing, storage, and communication technologies, as well as large amounts of software. It encompasses the many roles discussed in the Section 2, principal ones being the sharing of common resources, functions, and expertise among institutions and disciplines, as well as lowering the barriers to entry for the development, provisioning, operations and use of new applications.

From a budgetary perspective, there are significant challenges and opportunities that demand research. Significant advances are required in human-computer interaction, database systems, software engineering, networks, parallel computing, advanced architectures, security, reliability, interoperability and many other areas. While many present and future technologies can be acquired commercially and must be intelligently leveraged, these often do not meet the specialized needs of science and engineering research. Because these needs are often high-end and stretch available technologies, there is a significant opportunity to leverage the ACP to advance information technology itself, one of the important missions of NSF. Both the Internet and supercomputing architectures are historical illustrations of this process of turning the needs of academic researchers into valuable new technologies while simultaneously empowering the research community.

The cyberinfrastructure also raises numerous social issues, for example, those related to security, privacy, intellectual property, and use of information technology in support of research communities in collaborative work across distance, organizations, and disciplines, and associated new modes of scholarly communication. Research into these issues will also pay numerous dividends, both within the NSF community and in the nation as a whole.

Thus, the ACP requires significant basic research activities that address both the technical and social challenges as well as opportunities that surround the construction, management, and use of the nation's evolving cyberinfrastructure. The ACP must also evaluate the outcomes and support the evolution of the cyberinfrastructure to meet ever expanding needs.

Although a portion of these funds should support individual investigators exploring ground-breaking new activities, we also envision a number of larger multi-investigator projects that explore many technical and social issues and mixtures of the two, and involve substantial prototyping, testbeds, and experimentation. Each larger project needs substantial funding, averaging about $2 million annually. Past examples of this type of project include the Titanium[47] Compiler Project at UC-Berkeley, the Storage Resource Broker[48] project at SDSC, the DataCutter[49] project at Ohio State, and the Network Weather Service[50] Project at UC-Santa Barbara. This is also in line with large projects in the ITR program, which we view as successful in bringing together interdisciplinary teams addressing similar issues and we hope will continue as part of the base budget. We estimate conservatively that 30 to 40 projects would be needed to cover the breadth of research issues related to the proposed infrastructure, from making it usable to making it secure.

In our budget estimate we assume 30 projects, for a total of $60 million annually, spent largely on researchers, equipment, and supporting professional staff. Some appropriate and evolving level of these funds could be allocated to individual investigator grants keeping in mind that the CISE base budget will also support many such grants.

Research into the application of information technology to domain science and engineering research - The goal of the ACP is to revolutionize scientific and engineering research through the innovative application of the information technologies. While cyberinfrastructure is an important enabler for this to happen, the ACP also requires researchers within the domain-specific science and engineering research communities to collaborate with computer and information scientists and mathematicians and social scientists in identifying opportunities, refining these ideas through experiments and trials, and ultimately moving these application ideas into production, broad deployment and use. It also requires research into generic applications that span disciplines, and identification of common threads across applications that can be captured within the cyberinfrastructure.

This type of investigation allows research communities to take advantage of the new information technologies and infrastructure, as well as support development of new methods and facilities to tackle research challenges previously out of reach. This research will involve long-term efforts in the science and engineering disciplines, computer and information science, the social sciences, and mathematics. We

envision discipline scientists partnering with colleagues from other fields who can contribute to devising technical approaches to advance knowledge in new ways.

Once opportunities have been identified, they should be prototyped and introduced to real users, who will provide feedback to guide refinements and improvements. Ultimately, success will be measured by turning these applications into production software that is broadly adopted and used, and, importantly, many associated new processes and methodologies for the conduct of science and engineering research.

This activity should include a mixture of individual investigator and larger-scale grants or cooperative agreements. Turning successful prototypes into production, and the development of prototypes themselves, may call for partnerships with operational centers which offer expertise in software engineering, especially as these applications are turned over to production. On the other hand, one goal in advancing the cyberinfrastructure is to make it easier to develop and support new applications directly within application groups and disciplines. The distribution of grant sizes and types will likely vary by discipline. Successful models include the Grand Challenge awards of the mid 1990s and the application-oriented ITR grants of recent years. The large number of worthy but unfunded ITR proposals in recent years is a strong indicator of latent interest.

Our budget estimate is based on 50 grants at an average annual funding of $2 million, but also with considerable variation in grant size depending on discipline and the problem being tackled. Experience with application-oriented large ITR grants (roughly $2M-$3M/year for up to five years) has shown that some complex applications require substantially more funding. Some of these grants are large because of their interdisciplinary and inter-institutional character and the substantial needs for facilities, prototyping and experimentation, and supporting professional staff (software engineers, system administrators, user support, etc.).

Acquisition and development of cyberinfrastructure and applications - As the ACP evolves, increasing levels of support will be required for the development of production software, coupled with the licensing of commercial software components and the integration of the various custom and commercial components. Successful cyberinfrastructure and applications, as they move out of the prototype and experimentation stage, will require initial product creation, ongoing maintenance, upgrade, distribution, and user support. Where possible, any cyberinfrastructure and application software that is developed within this ACP should be subsequently commercialized, resulting in (hopefully) lower commercial licensing fees.

Cyberinfrastructure to support the myriad scientific and engineering applications will comprise many software tools, system software components, and other software building blocks. Examples of system software include grid middleware, parallelizing compilers for a variety of machine architectures, scalable parallel file systems and distributed databases, and sophisticated schedulers. Where appropriate these components will be commercially licensed, and NSF will purchase a "site" license on behalf of the community of NSF researchers.

An important activity will be an ongoing effort to identify the appropriate mix of commercial custom-developed software in accordance with an overall architectural plan, and then to acquire or develop and integrate these components. The outcome should be a single unified software distribution that users can download and install. Alternatively, centers will provision and operate this cyberinfrastructure and applications and offer them as services invoked over the network.

The NSF Middleware Initiative is exemplary of the type of program required to create and support the software aspects of cyberinfrastructure. While only a fraction of prototypes will require conversion to production status, the development costs of achieving the levels of stability and usability suitable for the larger community will require a development cost at least an order of magnitude greater than a prototype. The recurring costs of maintenance, upgrade, and user support will also be substantial. An active program to commercialize successful cyberinfrastructure and applications (especially the generic variety) will help to contain these costs.

These software development efforts would be supported wherever the expertise in computational science and software engineering is located, not just in large academic centers, and possibly in the commercial sector. Selection of the software development and maintenance groups should be based on expertise and experience, proposed plans and methodology, and anticipated costs. We estimate initially 20 such projects with an average annual budget of $5 million each.

We propose the creation and support of "cyberinfrastructure software centers" dedicated to developing the more difficult and sophisticated system and infrastructure software. These centers must have a scale necessary to attack the significant challenges of developing standards and production software for grids, programming tools, and data access and analysis, to name a few examples. Each center might employ on the order of 50 full-time-equivalent staff who would engage in professional software engineering, with a funding level in the $10M/year range. Ten such centers funded at this level would be a good starting point, with each center attacking one area, such as grid computing, compilers and runtime systems, visualization, program development environments, global scalable and parallel file systems, human computer interfaces, highly scalable operating systems, system management software, and so forth.

Provisioning and operations of cyberinfrastructure and applications - Whether software is acquired or developed, once it's integrated into a single distribution there are many operational issues to be addressed. Software to be downloaded and installed locally will need to be maintained (possibly) on multiple platforms, made available for download (including issues of authentication and access control), and supported through helpdesk facilities. Where services are provided over the network, the appropriate equipment and software must be acquired, integrated, installed, and operated, and again user support and helpdesk functions must be provided. While capital expenditures for facilities will be necessary, the bulk of the costs are recurring salaries for professional staff, including software engineers, system administrators and operators, and user support personnel. We anticipate that most of these activities will be conducted in centers funded under cooperative agreements with NSF.These needs can be broken down into several categories discussed in the following subsections.

High-end general-purpose centers

One class of centers will provide high-end computing resources, similar to the leading-edge sites of the current PACI program. These will feature some or all of the facilities currently found at such centers, including computers, large data archives, sophisticated visualization systems, collaboration services, licensed application packages, software libraries, digital libraries, very high-speed connections to a national research network backbone, and a cadre of skilled support personnel helping users take advantage of the facilities. Since the technologies deployed in these centers will be cutting-edge, the support staff also may have to develop software to provide missing functionality in the environment and to integrate the various resources and services.

Since progress in many science and engineering disciplines is paced by the capacity and peak performance of the available systems, as well as by the allocation and scheduling policies, there is need for both higher peak performance and higher capacity than currently available in the PACI program. The Panel strongly recommends the following principle: The United States academic research community should have access to the most powerful computers that can be built and operated in production mode at any point in time, rather than an order of magnitude less powerful, as has often been the case in the last decade.

The most powerful scientific computer in the world today is Japan's Earth Simulator System[20], with a peak speed of 40 teraflops (10^{12} floating point operations per second), built at a cost of around $400 million. DOE's Lawrence Livermore National Laboratory[51] (LLNL) has a 12 teraflops machine and its Los Alamos National Laboratory[52] is in the process of installing a 30 teraflops system. In FY 2004 (perhaps the first year of the ACP) at least one of the DOE laboratories is expected to install a system in the 60 – 100 teraflops range. All these systems

have been justified and are being used by a relatively small number of applications projects.

The Panel believes it is important that NSF make comparable systems available to the United States academic community, but, due to the large size and diversity of this community, such systems must support a much wider range of applications. If the U. S. academic community is to be competitive internationally in large-scale simulation, these considerations suggest systems in the 60 teraflops range in FY 2004, thereafter tracking the state of the art. In addition, at least a dozen individual American universities have acquired or are installing systems with peak speeds of over one teraflops. In order to enable new applications, national resources should be more powerful than those at individual universities by at least one to two orders of magnitude.

In terms of capacity, there should be a sufficient number of such systems that individual projects (with appropriate justification) can be granted the resource units to run many jobs per year that use a large fraction (at least 25%) of peak performance for tens or hundreds of hours. Such jobs usually access or produce vast amounts of data that need to be stored, visualized, and interacted with; hence, the entire environment needs to be balanced and scaled according to peak processing speeds. A typical balanced configuration meeting this criterion would have:

- At least 1 Byte of memory per FLOP/s.
- Memory Bandwidth (Byte/s/FLOP/s) ≥ 1.
- Internal Network Aggregate Link Bandwidth (Bytes/s/FLOP/s) ≥ 0.2.
- Internal Network Bi-Section Bandwidth (Bytes/s/FLOP/s) ≥ 0.1.
- System Sustained Productive Disk I/O Bandwidth (Byte/s/FLOP/s) ≥ 0.001.
- System High Speed External Network Interfaces (bit/s/FLOP/s) ≥ 0.00125.
- The internal network that connects the nodes with latency in the 1-2 microseconds or less, user memory to user memory.
- Globally addressable disks with at least 20 times the capacity of main memory.

Using those ratios, a 60 teraflops system with a balanced configuration would have 60 TB of memory and 1.2 PB of globally addressable disk space. Current estimates are that in FY 2004 such a system will cost on the order of $180 million. In FY 2007, $180 million might suffice to purchase a balanced system with a peak speed of 100 to 150 teraflops.

The panel recommends that about five such centers be supported; the two leading-edge sites of the PACI program plus the Pittsburgh Supercomputer Center should be considered as three of these centers, following appropriate review. While there are substantial economies of scale in operating large computers – a modestly larger staff can support a much larger computer or several systems – there are other

considerations in the number of centers. Each center tends to develop affinity with different disciplines or strengths in different aspects of information technology. Centers are training grounds for computational scientists and engineers, who then migrate to (more likely nearby) research institutions. A greater diversity of centers encourages novel approaches and new ideas. The primary measure of effectiveness of such centers is user satisfaction, and competition among a larger number of centers leads to greater satisfaction. On the other hand and as mentioned earlier, the number of centers is secondary and should in the end be based on additional analysis and community input.

There should be no shortage of institutions interested in creating and operating large-scale centers; over a dozen universities already operate substantial centers and have participated in previous competitions. For the purpose of the budget estimate we assume five centers, each with an annual budget of $75 million, for a combined annual budget of about $375 million ($300 million more than the current level). This is larger than the current centers primarily because we advocate higher-peak-performance and capacity computers and ancillary systems than at present. On the order of $50 million annually would be devoted to these equipment procurements, assuming that a major new system will be acquired by each center every three to four years. Most of the rest of the budget would be for recurring personnel costs; development, integration, maintenance, and upgrade of software; as well as provisioning and operations of cyberinfrastructure and user support.

In staging the operational portion of the ACP, in FY2004 and FY2005 (after appropriate review) the existing centers might acquire upgraded facilities and related infrastructure. (Spreading the ramp-up over two fiscal years will provide more choices and may increase performance as new generations of systems emerge). The second step might be to open a competition in FY2005 and FY2006 for additional centers.

Local clusters of computers are meritorious alternatives to centralized large systems for many needs, and in some research areas special-purpose hardware is the best option. The ACP will contribute to the creation of a grid environment (including middleware and tools) that will make all three options accessible to researchers at all institutions and facilitate the migration of applications from one to another. As with the number of centers, the balance of funding among these options should be based on additional analysis and community input.

A similar issue arises with professional support personnel. Budgets should seek (based on prior analysis) to achieve the best balance between local support (which can give more discipline-specific and intensive assistance) and centralized support (which benefits from economies of scale and scope and can usefully transfer expertise from one institution to another and from one discipline to another). A valuable middle ground is to locate discipline-specific groups at large centers.

Data repositories

Well curated data repositories are increasingly important to science and engineering research, allowing data gathered and created at great expense to be preserved over time and accessed by researchers around the world, including by disciples of other disciplines. The ACP should provide long-term and sustained support of such repositories. This involves much more than simply running large storage facilities. Supported by research into cyberinfrastructure, better ways to organize and manage such large repositories will be developed, and software infrastructure and tools will be developed, distributed, maintained, and supported. Appropriate standards will be developed that allow data to be self-documenting and discoverable through automated tools, and to insure the interoperability necessary to incorporate data acquired in one discipline into applications serving other disciplines.

To illustrate some detailed issues, data need to be organized in appropriate ways, metadata (machine readable and searchable descriptions of the data) must be systematically created, and basic manipulation and analysis tools provided. Data must be structured in ways that support both intra- and inter-discipline interoperability. Useful data repositories are also highly dynamic, requiring reclassification based on reanalysis of content. Migration of data to new media for preservation, and exploitation of higher capacity media is required. High-speed access to repositories by remote users raises capacity and scalability issues, with implications for their network, storage, and I/O subsystems.

As with computing, the cost of data repositories (done correctly) will be dominated by the recurring costs of personnel performing curation, maintenance and upgrade, and providing user advice, assistance, and support. The most sophisticated of these personnel need professional skills in the relevant aspects of information management and information technology (e.g., data bases, archival file systems, building portals), and will be developing and maintaining custom software. By using a combination of high-speed networks and local high-speed caches, there is no hard requirement to co-locate professional staff with physical storage particularly staff performing data acquisition and curation functions as opposed to disk partitioning, regeneration, and backup functions. As with computing, there is need for support personnel at local institutions, in discipline-specific groups (often located in centers), and centralized in centers. Although further analysis is needed, we expect that the most efficient approach will be to have relatively centralized storage hardware (with supporting staff) but distributed data acquisition and curation personnel. The balance of funding across these options should be determined by analysis and community input.

The challenge of data acquisition, curation, and access cannot be addressed solely by NSF, since other agencies in the United States

and internationally also support repositories. For example, NIH supports certain biology and biomedical data collections and NASA funds many archives of astronomy and remote sensing data. NSF should support repositories for a number of disciplines, such as astronomy, atmospheric and oceanic sciences, biology, biomedicine, climate modeling and observations, engineering of many variations, environmental and earth sciences, geophysics, high-energy physics, neuroscience, nuclear physics, and space sciences, among others. One can easily envision 50 to 100 such repositories. Indeed, a Web search quickly yields scores of existing repositories, many of which will not scale to future demands, interoperate well among disciplines, nor guarantee long-term access. Based on current experience, each repository will require $1.5 million to $3 million annually, not even including the substantial additional effort required to produce clean, well-documented data that retains long-term access and value. Overall these repositories may require $150 million annually, assuming 75 such repositories with an average yearly budget of $2 million. The number of physical locations for storage farms and supporting personnel may be considerably smaller than the number of disciplinary repositories, based on analysis of tradeoffs between community responsiveness and the availability of discipline-specific expertise vs. economies of scale and scope.

In addition, it is important to maintain ongoing development centers that address issues spanning all disciplines and ensure that the latest outcomes from the research community (including research funded under this ACP) and the commercial sector are applied to the expanding data storage and management challenge. These centers would be primarily responsible for spreading the latest technologies and best practices and insuring interoperability across disciplines through appropriate standardization. They are the primary point of connection to the computer and information sciences research communities (including the digital library, knowledge management, and knowledge mining communities) for the derivation, description, and management of the knowledge derived from computations and observational data. We recommend that approximately five such centers be established at an estimated cost of $3 million per year each, for a total of $15 million per year. In some cases these centers may be co-located with significant data repositories.

The Panel also recommends the creation of teams that would work on discipline-specific metadata standards, data formats, tools, access portals, etc., as well as help to select and install software, e.g., for the grid and databases. If one such effort is supported at $2 million per year for each of the ten disciplines listed above, a combined funding level of $20 million per year will be required.

Digital libraries

An integral component of cyberinfrastructure includes the nation's digital libraries, an area where NSF is already providing intellectual and organizational leadership. These libraries contain (much more so in the

future than today) our intellectual legacy, a fundamental resource for our scientific and engineering research and engineering practice.

NSF digital library initiatives have created new infrastructure and content of value to specific disciplines (including many in the humanities). It is important to continue such efforts through ongoing research, prototyping and experimentation with digital library technologies, development and deployment of proven solutions, and support for specific digital library repositories in disciplines represented at NSF. The potential has been barely tapped, and there is an opportunity to find and implement new mechanisms for sharing, annotating, reviewing, and disseminating knowledge. We suggest that the topic of digital libraries be broadened to consider even larger questions about the transformation of scholarly communication, including not only the accessing and sharing of knowledge, but also including this expanding knowledge as an integral element of the active collaboration among scholars.

The soon-to-conclude second phase of the NSF digital library initiative is investing about $10 million per year. Given the success of the initiatives, and the promise and critical importance of the area, we believe the budget should be at least $30 million per year for digital libraries activities, with a mix of project sizes from $1-3 million annually.

Networking

High-speed networks are a critical cyberinfrastructure facilitating access to the large, geographically distributed computing resources, data repositories, and digital libraries. As the commodity Internet is clearly not up to the task for high-end science and engineering applications, especially where there is a real-time element (e.g. remote instrumentation and collaboration), a high-speed research network backbone should be established and the current connections program extended to support access to this backbone as well as to provide international connections. Today we could aim for a 40 Gb/s (gigabit per second throughput) backbone with large center or user sites connecting at 10 – 40 Gb/s. Over time these numbers could increase rapidly with advances in technology and sustained funding. Assuming that 50 sites connect at 10 Gb/s and 40 sites at 40 Gb/s, a cost estimate of the backbone and connections is about $60 million per year.

As with computing, the primary issues in the backbone network are peak speed of data transfer and total capacity. The peak speed should be determined primarily by currently available production network equipment, and capacity upgrades will require ongoing monitoring and analysis to avoid significant congestion-induced communication latencies. However, from the perspective of applications and users, the performance of the backbone network is only one element of overall performance, which is also affected by local area networks, various processing and caching bottlenecks, processing delays in middleware and operating system layers, and computer I/O bandwidths, among others. For this reason, the research and development addressing

performance issues within the ACP should focus its attention on overall system performance, seeking out bottlenecks and removing those bottlenecks through research into underlying technology advances, system architectures (e.g. the strategic location of caching), and development of more advanced hardware and software solutions. The adequate funding of facilities upgrades and the funding of these research and development activities are equally important in providing the research community with state-of-the-art facilities.

Within this operations portion of the ACP, system measurement instruments and software should be deployed, a knowledge database of issues and solutions should be developed and maintained, and professional support staff should advise, assist and support researchers and applications developers encountering difficult performance issues.

A number of unique scientific facilities utilized by U.S. science and engineering communities are located outside the United States -- some even funded by the NSF (e.g. the Gemini South Observatory[53] in Chile). As noted elsewhere in this report, international collaboration is essential in research, and the United States has a vital interest in ensuring that its science and engineering community has high-speed access to the international infrastructure. NSF needs to connect the national backbone to similar infrastructure in other countries, and cooperate in other ways through research, development, standardization, and operations.

While these budget estimates may seem low, as throughout this section, this estimate is in addition to current NSF network research and infrastructure networking expenditures (as we have specified throughout this section), which is currently about $40 million annually for networking infrastructure. We also expect that individual states (e.g., California, Illinois, Indiana, and North Carolina) and individual universities will make coordinated investments to ensure that institutional infrastructure provide appropriate connectivity from the national backbone all the way to researchers' desktops.

Application service centers

In the budget categories already addressed, there are clearly some unmet needs, such as support services for non-computational applications, visualization, collaboration, or distributed and cluster computing, among others. These services may be provided through a combination of a utility model (making them available on the network) and by providing software distributions and support personnel to aid in their installation and use. As the research and development portions of the ACP yield successful outcomes, the needs in this area will expand. Initially the Panel recommends funding a modest number of centers to initiate these activities (five with an annual funding of $2 million each would be reasonable). Over time this budget (and the size and number of centers) would grow, guided by the successful models and expanding needs and opportunities.

6.4 Summary

The scope and scale of the ACP will require an annual budget of about $1 billion over and above the current PACI and network infrastructure programs in NSF.

We estimate that about 65% of the total budget is for the recurring costs of professional staff and researchers, as opposed to the acquisition of hardware and software. A substantial portion of these recurring costs is devoted to developing, maintaining, distributing, upgrading, and supporting software. This emphasis is consistent with past President's Information Technology Advisory Committee[33] (PITAC) recommendations for substantially greater investments in software research and production.

The implementation recommendations and budgets sketched in this chapter are based on experience with related projects and activities and reflect numerous comments and suggestions received from community leaders. Nevertheless, our recommendations should be considered as only a beginning. The ACP will require ongoing planning, implementation and adequate resources if it is to achieve its goal of revolutionizing the conduct of science and engineering research. All NSF directorates must participate in the planning and in the implementation in order to ensure that the cyberinfrastructure that is built is effective in bringing about this revolution.

7.0 References

1. NSF-CISE, http://www.cise.nsf.gov/
2. Partnership for Advanced Computational Infrastructure, http://www.paci.org/
3. National Center for Supercomputing Applications, http://www.ncsa.uiuc.edu/
4. San Diego Supercomputing Center, http://www.sdsc.edu/
5. Pittsburgh Supercomputing Center, http://www.psc.edu/
6. Teragrid, http://www.teragrid.org
7. NSF Middleware Initiative, http://www.nsf-middleware.org/NSF
8. NSF Digital Library Initiative, Phase 2, http://www.dli2.nsf.gov
9. George E. Brown, Jr. Network for Earthquake Engineering Simulation, http://www.nees.org/
10. The National Virtual Observatory, http://www.us-vo.org/
11. The National Ecological Observatory Network (NEON), http://www.nsf.gov/bio/neon/start.htm
12. The National Science Digital Library (NSDL), http://www.nsdl.nsf.gov/indexl.html
13. The Grid Physics Network (GriPhyN), http://www.griphyn.org/index.php
14. The Space Physics and Aeronomy Research Collaboratory (SPARC), http://intel.si.umich.edu/sparc/ and http://www.crew.umich.edu/
15. The Biomedical Informatics Research Network (BIRN), http://www.nbirn.net/
16. DOE National Collaboratories Program, http://doecollaboratory.pnl.gov/
17. DOE Scientific Discovery Through Advanced Computing (SciDAC), http://www.osti.gov/scidac/
18. UK Research Councils E-science Program, http://www.research-councils.ac.uk/escience/
19. European Commission Sixth Framework Research Program, http://europa.eu.int/comm/research/fp6/index_en.html
20. Japanese Earth Simulator Center, http://www.es.jamstec.go.jp/esc/eng/

21. Internet 2, UCAID, http://www.internet2.edu/
22. Global Grid Forum, http://www.gridforum.org/
23. Web Services Activity, http://www.w3.org/2002/ws/
24. *Preparing for the Revolution: Information Technology and the future of the Research University*, National Academy Press, 2002, http://www.nap.edu/catalog/10545.html
25. National Research Council, *Capacity of U.S. Climate Modeling to Support Climate Change Assessment Activities*, National Academy Press, Washington DC, 1998.
26. NSF Advisory Committee on Environmental Research and Education, http://www.nsf.gov/geo/ere/ereweb/advisory.cfm
27. Connecting European Research, GEANT, http://www.dante.net/geant/
28. National Center for Atmospheric Research, http://www.ncar.ucar.edu/ncar/
29. National Energy Research Scientific Computing Center, http://www.nersc.gov/
30. *Graduate Education for Computational Science and Engineering*, SIAM Working Group on CSE Education, available at http://www.siam.org/cse/report.htm
31. Society for Industrial and Applied Mathmatics (SIAM) website, http://www.siam.org/
32. Education, Outreach and Training Partnership in PACI, http://www/eot.org/
33. Advanced Network with Minority Serving Institutions Initiative (ANMSI), http://www.anmsi.org/
34. President's Information Technology (IT) Advisory Committee (PITAC), http://www/ccic.gov/ac/report/
35. Experimental Program to Stimulate Competitive Research (EPSCoR) http://www.ehr.nsf.gov/epscor/
36. Access Grid, http://www-fp.mcs.anl.gov/fl/accessgrid/
37. Report of the Task Force on the Future of the NSF Supercomputer Centers Program (Hayes Report), http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf9646
38. NSF Blue Ribbon Panel on High-Performance Computing (Branscomb Report), http://www.cise.nsf.gov/div/acir/hilderbrandt/branscomb/challenges.htm
39. Adapted from talk by Larry Smarr, *Science in the Connected World,* presented at AAAS Annual Meeting, Feb 26, 2002.
40. The Globus Project, http://www.globus.org/

41. Condor Project, http://www.cs.wisc.edu/condor/
42. Cactus Programming Framework, http://kevin.alteu.com/code/oldcode/cactus/tutorial.html
43. National Academies Press, http://www.nap.edu/books/0309070317/html/
44. GridLab, http://www.gridlab.org/
45. Terasale Computing System, http://www.psc.edu/publicinfo/terascale/bigiron.html
46. Information Technology Research (ITR), http://www.itr.nsf.gov/
47. Titanium Compiler Poject, http://www.cs.berkeley.edu/Research/Projects/titanium/
48. SDSC Storage Resource Broker, http://www.npaci.edu/DICE/SRB/
49. DataCutter Project, http://medicine.osu.edu/informatics/DIGC/dc/
50. Network Weather Service Project, http://nws.cs.ucsb.edu/
51. Lawrence Livermore National Laboratory (LLNL), http://www.llnl.gov/
52. Los Alamos National Laboratory, http://www.lanl.gov/
53. Gemini South Observatory, http://www.gemini.edu/